ESTIMATING THE PROBABILITY OF BEING THE BEST SYSTEM: A GENERALIZED METHOD AND NONPARAMETRIC HYPOTHESIS TEST

THESIS

Aaron M. Lessin, Captain, USAF

AFIT-ENS-13-M-10

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT-ENS-13-M-10

ESTIMATING THE PROBABILITY OF BEING THE BEST SYSTEM: A GENERALIZED METHOD AND NONPARAMETRIC HYPOTHESIS TEST

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Aaron M. Lessin, BS

Captain, USAF

March 2013

AFIT-ENS-13-M-10

ESTIMATING THE PROBABILITY OF BEING THE BEST SYSTEM:
A GENERALIZED METHOD AND NONPARAMETRIC HYPOTHESIS TEST

Aaron M. Lessin, BS
Captain, USAF

Approved:

//Signed//	21 Feb 13
Dr. J. O. Miller (Chairman)	Date
//Signed//	21 Feb 13
Dr. Raymond R. Hill (Member)	Date

AFIT-ENS-13-M-10

Abstract

This thesis provides two new approaches for comparing competing systems. Instead of making comparisons based on long run averages or mean performance, the first paper presents a generalized method for calculating the probability that a single system is the best among all systems in a single trial. Unlike current empirical methods, the generalized method calculates the exact multinomial probability that a single system is best among competing systems. The ability to avoid time consuming empirical estimation techniques could potentially result in significant savings in both time and money when comparing alternate systems. A Monte Carlo simulation study is conducted comparing the empirical probability estimates of the generalized integral method, calculated using density estimation technique, with those of two related estimation techniques, Procedure BEM (Bechhofer, Elmaghraby, and Morse) [1] and Procedure AVC (All Vector Comparisons) [2]. All test cases show comparable performance in empirical estimation accuracy of the generalized integral method with that of the current methods analyzed.

The second paper proposes the use of a distribution-free ordered comparisons procedure to test whether one population is stochastically larger (or smaller) than the other. This procedure is suggested as a useful alternative to the well-known Wilcoxon rank sum and Mann-Whitney tests. A Monte Carlo study is conducted, comparing the methods based on simulated power and type I error. All test cases show a marked improvement in performance of the ordered comparisons test versus the Wilcoxon rank sum test, when testing for stochastic dominance. A table of critical values is presented and a large sample approximation is given.

In Loving Memory of Laurian Leiker

Acknowledgements

Although I am credited with the composition of this thesis, this work would not have been possible without the contributions and support of many others. First, I need to thank my beautiful wife for her unending support and love throughout my research and time at AFIT. She was always there to pick up my slack during the countless long days and nights I spent working on my studies, all while taking care of our three wonderful children.

I am also indebted to my advisor, Dr. J.O. Miller, for his willingness to take me on as an advisee and allow me to pursue my mathematical interests, wherever they might lead. His constant guidance and support made this effort a truly enjoyable experience.

Additionally, I would like to thank Dr. Raymond Hill for always having an open door and an endless supply of literature resources. His comments and suggestions to this work were invaluable.

Finally, on behalf of all my peers in GOR13M, I would like to give all our instructors at AFIT the acknowledgment they deserve for their selfless mentorship and dedication to making us not only better analysts but officers as well.

Aaron M. Lessin

Table of Contents

Page

List of Figures

Page

List of Tables

ESTIMATING THE PROBABILITY OF BEING THE BEST SYSTEM: A GENERALIZED METHOD AND NONPARAMETRIC HYPOTHESIS TEST

I. Introduction

1.1 Overview

Suppose we have $k \geq 2$ independent populations, denoted as $\pi_1, \pi_2, \dots \pi_k$ where the goal is to determine which of the k systems is best. It is common practice when comparing competing systems to simply call system i the "best" if it has the most desirable mean or long-run average performance. However, in situations where we have only a small sample of data or where one-shot performance is important (i.e. building a satellite or firing a missile), we may be unable to compare the systems based on long-run performance. Instead, our goal is to find the system that is most likely to perform best in a single trial among all systems; this is referred to as the multinomial selection problem. Specifically, we will focus our attention on the problem of estimating the probability that each system will be the best in a single comparison among all the systems.

In addition to multinomial probability estimation techniques, many nonparametric procedures exist for the two-sample comparison problem, such as the commonly used Wilcoxon rank sum [3] and Mann-Whitney [4] tests. However, in terms of detecting whether one population is stochastically larger (or smaller) than the other, it appears that these test procedures may not make full use of all the information contained in the sample data. We present a new distribution-free hypothesis test based on ordered comparisons which gleans useful information about the empirical cumulative probability distributions of the sample data by ordering each sample separately (as opposed to a combined ordering as in the Wilcoxon rank sum test). The ordered sets of observations are then compared and a test statistic is calculated. Essentially, if sample 1 has significantly more smaller values than sample 2, then, in general, the cumulative probability distribution of sample 1 will be monotonically greater than the cumulative probability distribution of sample 2. This will result in population 2 being stochastically larger than population 1.

Throughout this thesis, a hypothetical example will be considered from which example calculations will be conducted on a simulated dataset in order to motivate the work therein. Specifically, imagine the Air Force has recently experienced an unusually high number of critical failures on a specific Air Force system (i.e. an aircraft or computer system). As a result, the Air Force would like to implement a preventative maintenance policy to help thwart these failures and extend the time between critical failures. Two preventative policies are being considered for implementation (we will refer to them simply as Policy A and Policy B). A computer simulation was conducted using each of the policies, and output data was collected on the resulting time between critical failures (see Table 1) experienced using each policy. This data will be used throughout the thesis to demonstrate not only existing analysis techniques, but the proposed generalized integral method for calculating the probability that each system will be the best in a single comparison among all the systems as well as conducting the nonparametric ordered comparisons hypothesis test. At the conclusion of the thesis, we will provide a recommendation as to which policy the Air Force should implement in order to extend the time between critical failures.

1.2 Thesis Organization

This thesis is formatted as two separate papers. Chapter 2 consists of the first paper which presents a generalized method for calculating the probability that each system is best in a single comparison among all systems. An overview of the relevant literature is provided, followed by an introduction to the generalized method. A Monte Carlo simulation study is also conducted, comparing the proposed method with two closely related techniques.

The second paper, in Chapter 3, proposes a nonparametric ordered comparisons hypothesis test. A definition of nonparametric statistics, the distribution-free property, and stochastic dominance is given, followed by a brief summary of the Wilcoxon rank sum and Mann-Whitney tests. The ordered comparisons test procedure is presented, along with a derivation of the associated critical values. A large-sample approximation is also formulated, followed by a numerical example demonstrating the ordered comparisons technique. Finally, we compare the ordered comparisons test to the Wilcoxon rank sum test based on the simulated power and type I error, via a Monte Carlo simulation study.

II. A Generalized Method for Estimating the Probability of Being the Best System

2.1 Abstract

This paper presents an alternate method for comparing competing systems. Instead of making comparisons based on long run averages or mean performance, a generalized method is proposed for calculating the probability that a single system is the best among all systems in a single trial. Unlike current empirical methods, the generalized method calculates the exact multinomial probability that a single system is best among competing systems. The ability to avoid time consuming empirical estimation techniques could potentially result in significant savings in both time and money when comparing alternate systems. A Monte Carlo simulation study is conducted comparing the empirical probability estimates of the generalized integral method, calculated using a density estimation technique, with those of two related estimation techniques, Procedure BEM (Bechhofer, Elmaghraby, and Morse) [1] and Procedure AVC (All Vector Comparisons) [2]. All test cases show comparable performance in empirical estimation accuracy of the generalized integral method with that of the current methods analyzed.

2.2 Introduction

Suppose we have $k \geq 2$ independent populations, denoted as $\pi_1, \pi_2, \dots \pi_k$ where the goal is to determine which of the k systems is best. It is common practice when comparing competing systems to simply call a system the "best" if it has the most desirable mean or long-run average performance. However, in situations where there is only a small sample of data or where one-shot performance is important (i.e. building a satellite or firing a missile), we may be unable or unwilling to compare the systems based on long-run performance. Instead, it may be more beneficial to find the system that is most likely to perform best in a single trial among all systems; this is referred to as the multinomial selection problem. Specifically, we will focus our attention on the problem of estimating the probability that a simulated system is best in a single comparison among all systems.

Section 2.3 of this paper provides a brief overview of the two most relevant multinomial selection procedures, Procedure BEM [1] and Procedure AVC [2], from which we develop and propose a generalized method. Section 2.4 introduces this generalized method for calculating the probability that each system is the best among competing systems in a single trial. Section 2.5 compares our method with the BEM and AVC procedures through a Monte Carlo simulation study.

2.3 Background

Traditionally, hypothesis tests of homogeneity are conducted by comparing a parameter of interest (typically the mean) of k systems under the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

where μ_i is the parameter of interest for the i^{th} system. However, when the null hypothesis is rejected, we are limited in our ability to determine which system (or systems) is in fact better (or worse) than the others. Ranking and selection techniques give us the capability to determine these differences between competing systems.

Ranking and selection methods provide the ability to order a set of alternative systems from worst to best, select the single best system from among a relatively small number of systems, or screen a large number of systems by removing those systems whose performance is statistically inferior. These techniques trace their roots to three main papers; Gupta [5] [6] pioneered a screening subset selection formulation for selecting a subset of alternatives which contains the best system, and Bechhofer [7] established the indifference zone formulation for selecting the single best system.

Suppose we have k systems of interest. Let X_{ji} be the i^{th} replication of the performance parameter of interest from system j, where $j = 1, 2, \dots, k$. We assume that the X_{ji} observations are independent within and across the systems, with means $\mu_j = E\left[X_{ji}\right]$ and variances $\sigma_j^2 = \text{Var}\left[X_{ji}\right]$. Also, let

$$p_j = \Pr\left\{X_{ji} > X_{\ell i}, \qquad \forall \ell \neq j\right\} \tag{1}$$

be the probability that system j will have the most desirable value of the performance parameter on a single replication and $\sum_{j=1}^{k} p_j = 1$.

For our discussion on ranking and selection techniques, the alternative systems are compared based on their expected performance, μ_j, or their probability, p_j, of being the best system in a single trial. We assume that larger is better in both cases and let $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{k-1} \leq \mu_k$ and $p_1 \leq p_2 \leq \cdots \leq p_{k-1} \leq p_k$ so that (unknown to us) system k is best.

2.3.1 Subset Selection Formulation

Consider the case where the goal is to select a subset of the k systems which contains the best system. Since we can't be absolutely sure that this subset contains the best system, we specify a probability, P^*, that the subset contains the best system where $1/k < P^* < 1$. More formally, if we have v observations or simulation replications from each of the k systems, our goal is to obtain a subset $I \subseteq \{1, 2, \ldots, k\}$ such that

$$\Pr\{k \in I\} \geq P^*. \tag{2}$$

Ideally, we would like $|I|$ to be as small as possible; the best case scenario being $|I| = 1$ when the subset contains only the best system. Gupta [5] [6] first proposed a procedure for cases where the data is normally distributed with common variance σ^2 and we have the same number of observations from each system. Nelson et al. [8] developed additional techniques to handle a variety of scenarios, including those where the unknown variances are not equal.

Gupta's technique includes in the subset I all of the systems ℓ such that

$$\bar{X}_\ell(n) \geq \max_{j \neq \ell} \bar{X}_j(n) - h\sigma\sqrt{\frac{2}{v}} \tag{3}$$

where $\bar{X}_j(n)$ is the sample mean of the first v outputs from system j, and h is a constant whose value depend on k and P^* [9].

The proof that rule (3) guarantees (2) and shows what the value of h should be is given by Kim and Nelson [9] as follows:

$$\Pr\{k \in I\} = \Pr\left\{\bar{X}_k(n) \geq \max_{j \neq k} \bar{X}_j(n) - h\sigma\sqrt{\frac{2}{v}}\right\}$$

$$= \Pr\left\{\bar{X}_k(n) \geq \bar{X}_j(n) - h\sigma\sqrt{\frac{2}{v}}, \forall j \neq k\right\}$$

$$= \Pr\left\{\frac{\bar{X}_j(n) - \bar{X}_k(n) - \left(\mu_j - \mu_k\right)}{\sigma\sqrt{2/v}} \leq h - \frac{\left(\mu_j - \mu_k\right)}{\sigma\sqrt{2/v}}, \forall j \neq k\right\}$$

$$\geq \Pr\left\{Z_j \leq h, i = 1, 2, \ldots, k-1\right\} = P^*$$

where $(Z_1,\ Z_2, \ldots, Z_{k-1})$ have a multivariate normal distribution with means 0, variances 1, and common pairwise correlations 1/2.

Additional methods have also been developed which further Gupta's initial groundwork. For example, say we have a large number of competing systems to begin our analysis and we only have enough resources available to analyze a specific number of alternatives. Koenig and Law [10] developed a screening subset selection procedure where we specify a subset of exactly m systems which contain the best system with probability P^*. Gupta and Santner [11] extended the original method by specifying a maximum size m of the subset I. The benefit of defining m as the maximum subset size as opposed to ensuring that exactly m systems be selected, is that it is possible for far fewer than m systems to be selected when it is statistically clear that only a few of the systems could be the best [12].

2.3.2 Indifference Zone Formulation

A major disadvantage of the subset selection formulation is that it is possible, and likely, that the subset I will contain more systems than just the single best system. However, there is no subset selection technique that can guarantee a subset of size one and satisfy (2) for an arbitrary sample size. Ultimately, we would like to determine the single best system among all the competing systems. Bechhofer's [7] indifference zone formulation was developed to solve this problem.

The goal of the indifference zone formulation is to select the single best of k systems with specified probability, P^*, whenever $\mu_k - \mu_{k-1} \geq \delta$, where $\delta > 0$ is the smallest difference the experimenter deems worth detecting. This specified distance, δ, is known as the indifference parameter. Hence, if there are systems whose performance is within δ of the best, then we are indifferent as to which system is chosen as the best. Similar to the subset selection formulation, Bechhofer's technique guarantees that

$$\Pr\{\text{select system } k \mid \mu_k - \mu_{k-1} \geq \delta\} \geq P^* \tag{4}$$

where $1/k < P^* < 1$. To perform Bechhofer's procedure, begin by collecting

$$n = \left\lceil \frac{2h^2\sigma^2}{\delta^2} \right\rceil \tag{5}$$

observations from each system, where h is a constant (as determined below) and $\lceil\cdot\rceil$ means to round up; then select as best the system with the largest sample mean [9]. Assuming that $\mu_k - \mu_{k-1} \geq \delta$, Kim and Nelson [9] show that

$$\begin{aligned}
\Pr\{\text{select system } k\} &= \Pr\left\{\bar{X}_k(n) > \bar{X}_j(n), \forall j \neq k\right\} \\
&= \Pr\left\{\frac{\bar{X}_j(n) - \bar{X}_k(n) - \left(\mu_j - \mu_k\right)}{\sigma\sqrt{2/v}} < -\frac{\left(\mu_j - \mu_k\right)}{\sigma\sqrt{2/v}}, \forall j \neq k\right\} \\
&\geq \Pr\left\{\frac{\bar{X}_j(n) - \bar{X}_k(n) - \left(\mu_j - \mu_k\right)}{\sigma\sqrt{2/v}} < \frac{\delta}{\sigma\sqrt{2/v}}, \forall j \neq k\right\} \\
&\geq \Pr\left\{\frac{\bar{X}_j(n) - \bar{X}_k(n) - \left(\mu_j - \mu_k\right)}{\sigma\sqrt{2/v}} < h, \forall j \neq k\right\} \\
&\geq \Pr\left\{Z_j < h, i = 1, 2, \ldots, k-1\right\} = P^*
\end{aligned}$$

where again $(Z_1,\ Z_2, \ldots, Z_{k-1})$ have a multivariate normal distribution with means 0, variances 1, and common pairwise correlations 1/2.

Goldsman et al. [13] note that Bechhofer's procedure is essentially a power calculation that determines how many observations need to be collected in order to detect differences between competing systems of at least δ. In fact, when the differences in performance between systems are greater than δ, the number of observations, n, that we collect from each system may in fact be much larger than needed. Therefore, additional methods have been developed which collect observations and make decisions sequentially. This allows us to omit inferior systems much earlier in the process. Sequential techniques trace their roots to Wald [14], and the first procedure specifically formulated for sequential selection is credited to Paulson [15]. Law [12] provides an overview of additional indifference zone techniques.

2.3.3 Procedure BEM

Up to this point, we have focused our discussion on comparisons based on the expected performance, μ_j, of each system. However, ranking systems based on long-run mean performance may not always give us the full picture as to which system is best. It is possible for the system with the best mean performance to not be the system with the highest probability of being the best. Consider the following example presented by Goldsman [16]: Let A and B be two inventory policies, where profit is considered the parameter of interest and the higher the profit, the better. Suppose the profit from policy A is 1000 and occurs with probability 0.001 or 0 with probability 0.999. The profit from policy B is 0.999 with probability 1. The expected profit from policy A is 1 and the expected profit from policy B is 0.999. Therefore, we see that policy A has a higher expected profit then policy B. However, there is also a 0.999 probability that the profit from policy B is greater than the profit from policy A. Therefore, policy B almost always has a higher profit than policy A, even though the long-run expected profit for policy A is greater. Hence, it may be wise to consider policy B as the best policy. This simple example illustrates the need for an alternate method such as the probability estimation procedures presented in this paper when comparing competing systems.

Let p_j be the probability that system j has the most desirable value of the performance parameter on any replication containing one observation from each system as defined in (1). We assume that our performance parameter is a continuous random variable, and so the probability that a tie occurs between systems is exactly zero. Therefore, in each replication only one system can be the best; this corresponds to a multinomial trial where one and only one system can ‘win’ in any given trial [17]. If we have v independent replications, the number of wins for each system follows a multinomial distribution. The objective in the methods presented in the remainder of this paper is to estimate these unknown multinomial success probabilities, p_j, for each of the k systems [17].

Bechhofer and Sobel [18] made use of multinomial selection procedures to find the system most likely to produce the largest observation in a single trial. Bechhofer, Elmaghraby, and Morse [1] then developed the standard BEM procedure which determines the best system as the one having the largest value of the performance measure of interest in more trials than any other system. More formally, let

$$Y_{ji} = \begin{cases} 1, & \text{if } X_{ji} > X_{\ell i} \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, v$ and $\ell = 1, 2, \ldots, k$ but $\ell \neq j$ [17]. That is, $Y_{ji} = 1$ if X_{ji} is the largest observation among all systems in replication i. Then count the number of times that system j is the best across all of the v independent replications. Specifically,

$$Y_j = \sum_{i=1}^{v} Y_{ji} \tag{6}$$

for $j = 1, 2, \dots, k$. Then let $Y_{[1]} \leq Y_{[2]} \leq \cdots \leq Y_{[k]}$ be the ranked counts. Therefore, the system with the largest count, $Y_{[k]}$, is determined to be the best system [1]. If there is a tie for the largest count, randomly select one of the tied systems as the best. Then the estimate for the multinomial success probability, p_j, of each system is just

$$\hat{p}_j = \frac{Y_j}{v} \tag{7}$$

for $j = 1, 2, \dots, k$.

Consider the following simulated data in Table 1, which represents the time between critical failures (in months) for the two preventative maintenance policies being considered by the Air Force as mentioned during the thesis introduction.

Table 1. Simulated Time Between Critical Failures (In Months) for the Air Force's Alternative Preventative Maintenance Policies

Policy A	Policy B
0.31950	0.48383
0.76029	1.18579
0.26898	0.83970
1.40037	0.87125
3.49909	0.47615
0.01253	2.19601
0.27408	0.62893
0.10418	4.42322
1.55529	0.52177
0.81182	1.03523

Table 2 shows the Y_{ji} calculations and overall Y_j totals for both policies for the associated example data in Table 1.

Table 2. Y_j Example Calculations

Policy A	Policy B	Y_{1i}	Y_{2i}
0.31950	0.48383	0	1
0.76029	1.18579	0	1
0.26898	0.83970	0	1
1.40037	0.87125	1	0
3.49909	0.47615	1	0
0.01253	2.19601	0	1
0.27408	0.62893	0	1
0.10418	4.42322	0	1
1.55529	0.52177	1	0
0.81182	1.03523	0	1
		$Y_1 = 3$	$Y_2 = 7$

The multinomial probability estimate calculations for $\hat{p}_1$ and $\hat{p}_2$ are then

$$\hat{p}_1 = \frac{Y_1}{v} = \frac{3}{10} = 0.3$$

and

$$\hat{p}_2 = \frac{Y_2}{v} = \frac{7}{10} = 0.7$$

This same example data is used throughout this paper in order to provide a common frame of reference for demonstrating each procedure and providing an initial performance comparison between the different methods.

2.3.4 Procedure AVC

Building on the work of Bechhofer, Elmaghraby, and Morse [1], Miller, Nelson, and Reilly [2] developed an improved method known as Procedure AVC (All Vector Comparisons). In their method, the same v independent replications across k systems in Procedure BEM are used to form v^k pseudo-replications that contain one observation from each system [2]. Instead of only comparing the i^{th} replication for each system with the i^{th} replication from the other systems, Procedure AVC compares each X_{ji} $(j = 1, 2, \ldots, k;\ i = 1, 2, \ldots, v)$ with all possible combinations of the remaining observations, $X_{\ell h}$ $(\ell = 1, 2, \ldots, k;\ \ell \neq j;\ h = 1, 2, \ldots, v)$ [17]. Therefore, Procedure AVC makes use of $v^k - v$ additional comparisons than Procedure BEM. By using this additional information already contained in the sample data, Miller, Nelson, and Reilly are able to calculate more accurate estimates of the multinomial probabilities, p_j, for each system.

Specifically, let

$$Z_j = \sum_{a_1=1}^{v} \sum_{a_2=1}^{v} \cdots \sum_{a_k=1}^{v} \prod_{\ell \neq 1; \ell \neq j}^{k} \phi\left(X_{j a_j} - X_{\ell a_\ell}\right) \tag{8}$$

for $j = 1, 2, \dots, k$ with

$$\phi(a) = \begin{cases} 1, & a > 0 \\ 0, & a < 0 \end{cases}$$

as defined by Miller, Nelson, and Reilly [17]. Therefore, Z_j is simply the number of times that system j is the best out of the v^k pseudo-replications. Then let $Z_{[1]} \leq Z_{[2]} \leq \cdots \leq Z_{[k]}$ be the ranked counts. Again, the system with the largest count, $Z_{[k]}$, is determined to be the best system [2]. If there is a tie for the largest count, randomly select one of the tied systems as the best. Then the estimate for the multinomial success probability, p_j, of each system is just

$$\bar{p}_j = \frac{Z_j}{v^k} \tag{9}$$

for $j = 1, 2, \dots, k$.

We will again look at an example calculation using the same data as previously shown in Table 1. Table 3 shows the Z_j totals for both policies for the associated aforementioned data. Instead of having only the 10 original replications, we now have 100 pseudo-replications for comparing the two preventative maintenance policies.

Table 3. Z_j Example Calculations

Policy A	Policy B	X_{1i}	X_{2h}
0.31950	0.48383	0	1
0.31950	1.18579	0	1
0.31950	0.83970	0	1
0.31950	0.87125	0	1
0.31950	0.47615	0	1
0.31950	2.19601	0	1
0.31950	0.62893	0	1
0.31950	4.42322	0	1
0.31950	0.52177	0	1
0.31950	1.03523	0	1
⋮	⋮	⋮	⋮
0.81182	0.48383	1	0
0.81182	1.18579	0	1
0.81182	0.83970	0	1
0.81182	0.87125	0	1
0.81182	0.47615	1	0
0.81182	2.19601	0	1
0.81182	0.62893	1	0
0.81182	4.42322	0	1
0.81182	0.52177	1	0
0.81182	1.03523	0	1
		$Z_1 = 33$	$Z_2 = 67$

The multinomial probability estimate calculations for $\bar{p}_1$ and $\bar{p}_2$ are then

$$\bar{p}_1 = \frac{Z_1}{v^k} = \frac{33}{10^2} = 0.33$$

and

$$\bar{p}_2 = \frac{Z_2}{v^k} = \frac{67}{10^2} = 0.67$$

We can see how there are significantly more comparisons made in Procedure AVC than in Procedure BEM, which ultimately leads to slightly different estimates for the probability that each population is the best. Next, we'll show how our new method builds off the work of Bechhofer, Elmaghraby, and Morse [1] as well as Miller, Nelson, and Reilly [2], but approaches the problem in a more generalized manner. In general, our approach estimates the probability density functions or the cumulative density functions for each system and then evaluates a joint density function to come up with our multinomial probability estimates.

2.4 Generalized Integral Method

In order to develop the generalized integral method, let's begin by taking a different look at the example data from Table 1. Histograms of the data for Policy A and Policy B are shown below in Figure 1 and Figure 2, respectively. We arbitrarily use a common bin width of 0.75 for both histograms in order to illustrate initial calculations using our method. Different bin sizes may provide different results, with the most accurate estimates obtained from using individual observations from each system.

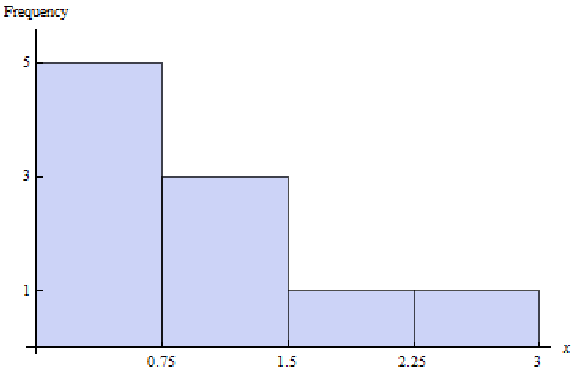

Figure 1. Histogram of Policy A Example Data

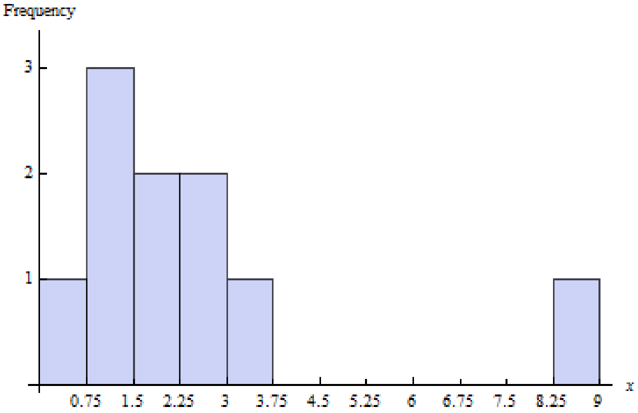

Figure 2. Histogram of Policy B Example Data

Likewise, we'll also consider the empirical cumulative distribution plots for each policy; these graphs are presented in Figure 3 and Figure 4 below, again using an arbitrary common bin width of 0.75.

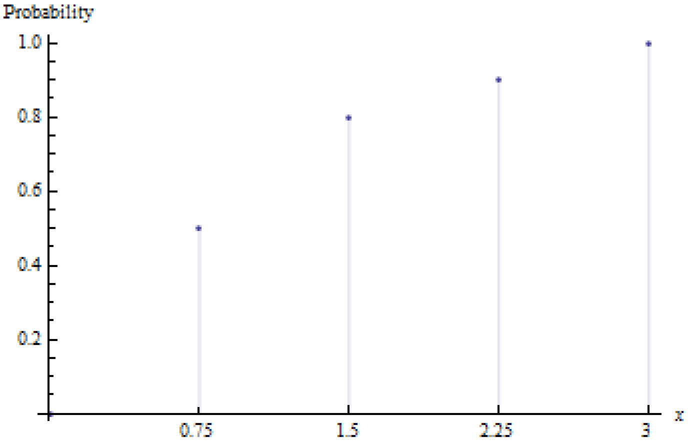

Figure 3. Empirical Cumulative Distribution Plot of Policy A Example Data

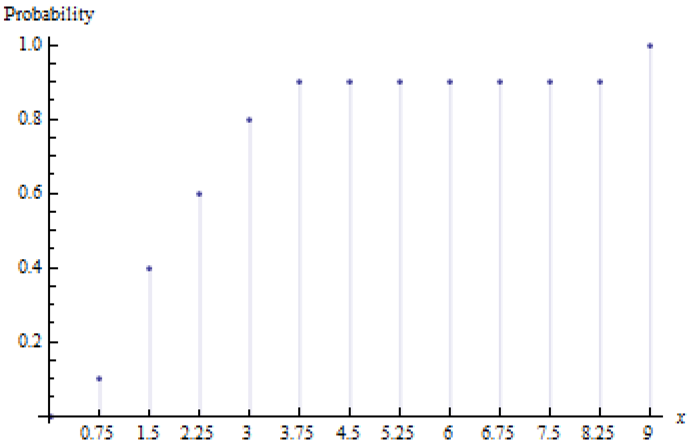

Figure 4. Empirical Cumulative Distribution Plot of Policy B Example Data

When estimating the multinomial probability, p_j, system j is considered best when, for any value of x from system j, all other systems perform worse than x. Therefore, what we are truly concerned with is the probability of x occurring in system j and the probability of any value less than or equal to x occurring in all other systems.

For two preventative maintenance policies, we can calculate the probability of Policy A having a larger time between critical failures than Policy B by multiplying the probability of obtaining a value X_1, using the probability density function (pdf) of Policy A, by the probability of obtaining a value X_2 less than X_1 using the cumulative distribution function (cdf) of Policy B. To further illustrate using our empirical pdf for Policy A from Figure 1 and our empirical cdf for Policy B from Figure 4, we can estimate the overall probability that Policy A is the best in a single trial as

$$\begin{aligned}\tilde{p}_1 &= (\Pr\{0 < X_1 \le 0.75\})(\Pr\{X_2 \le 0.75\}) \\ &\quad +(\Pr\{0.75 < X_1 \le 1.5\})(\Pr\{X_2 \le 1.5\}) \\ &\quad +(\Pr\{1.5 < X_1 \le 2.25\})(\Pr\{X_2 \le 2.25\}) \\ &\quad +(\Pr\{2.25 < X_1 \le 3.0\})(\Pr\{X_2 \le 3.0\}) \\ &= (0.5)(0.1) + (0.3)(0.4) + (0.1)(0.6) + (0.1)(0.8) \\ &= 0.31\end{aligned}$$

and by the same method we find that $\tilde{p}_2 = 0.69$.

As we increase the sample size from each simulated maintenance policy, eventually we can move from using the empirical pdfs and cdfs for the time between critical failures of each policy to using theoretical pdfs and cdfs for each policy. In fact, the data used in Table 1 were randomly selected from two exponential distributions where $\mu_1 = 1$ $(\lambda_1 = 1)$ and $\mu_2 = 2$ $(\lambda_2 = 1/2)$. Figure 5 shows the theoretical pdfs used to generate the example data for both policies.

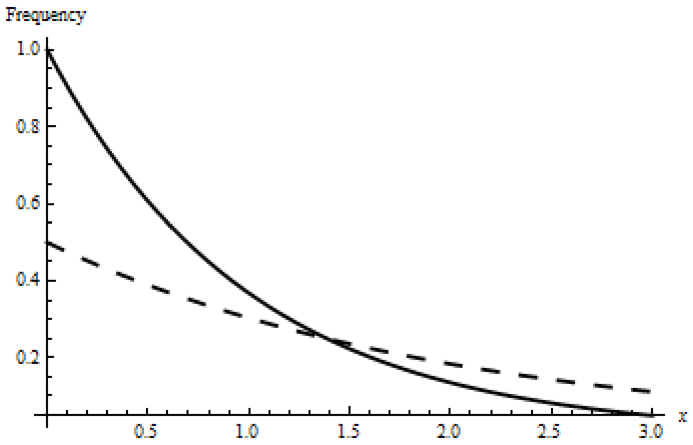

Figure 5. Probability Density Functions for Policy A (Solid Curve) and Policy B (Dashed Curve)

Figure 6 shows the theoretical cdfs for both preventative maintenance policies.

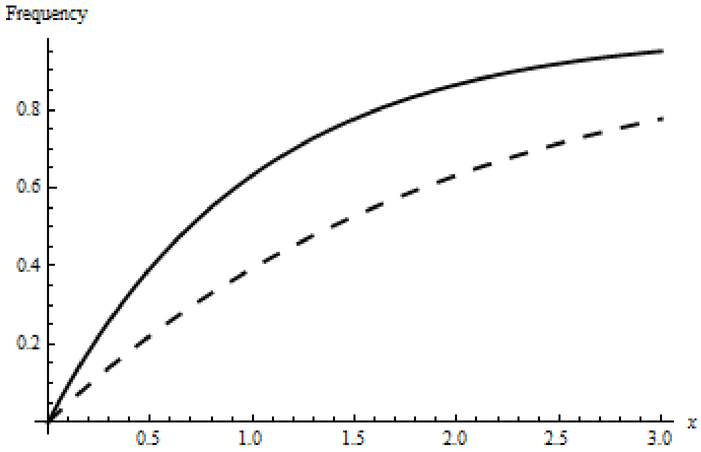

Figure 6. Cumulative Distribution Functions for Policy A (Solid Curve) and Policy B (Dashed Curve)

Following our previous multinomial probability calculations using empirical distributions, we can now substitute the theoretical distributions shown in Figure 5 and Figure 6. Let

$$f_1(x) = e^{-x}$$

and

$$f_2(x) = (1/2)e^{-(1/2)x}$$

be the pdfs for Policy A and Policy B, respectively.

Also, let

$$F_1(x) = 1 - e^{-x}$$

and

$$F_2(x) = 1 - e^{-(1/2)x}$$

be the cdfs for Policy A and Policy B, respectively.

We can calculate the multinomial probabilities for each policy in the same manner as we did using the empirical distributions. We have

$$\begin{aligned} \tilde{p}_1 &= \int_0^\infty f_1(x)F_2(x)dx \\ &= \int_0^\infty (e^{-x})\left(1 - e^{-(1/2)x}\right)dx \\ &= \frac{1}{3} \end{aligned} \tag{10}$$

and

$$\begin{aligned} \tilde{p}_2 &= \int_0^\infty f_2(x)F_1(x)dx \\ &= \int_0^\infty \left((1/2)e^{-(1/2)x}\right)(1 - e^{-x})dx \\ &= \frac{2}{3} \end{aligned} \tag{11}$$

We can see that these multinomial probabilities are similar to the probability estimates found using Procedure BEM ($\hat{p}_1 = .3,\ \hat{p}_2 = .7$) and Procedure AVC ($\bar{p}_1 = .33$, $\bar{p}_2 = .67$). In fact, if we use the individual observations to construct the empirical pdfs and cdfs for the example data in Table 1 as opposed to using an arbitrarily assigned band width, we get the exact same results as using Procedure AVC. Specifically, we can calculate the multinomial probabilities for each policy as

$$\begin{aligned}
\tilde{p}_1 &= \left(\Pr\left\{0 < X_1 \le x_{1(1)}\right\}\right)\left(\Pr\left\{X_2 \le x_{1(1)}\right\}\right) + \\
&\quad \vdots \\
&\quad +\left(\Pr\left\{x_{1(9)} < X_1 \le x_{1(10)}\right\}\right)\left(\Pr\left\{X_2 \le x_{1(10)}\right\}\right) \\
&= (\Pr\{0 < X_1 \le 0.01253\})(\Pr\{X_2 \le 0.01253\}) + \\
&\quad \vdots \\
&\quad +(\Pr\{1.55529 < X_1 \le 3.49909\})(\Pr\{X_2 \le 3.49909\}) \\
&= (0.1)(0) + (0.1)(0) + (0.1)(0) + (0.1)(0) + (0.1)(0) + (0.1)(0.4) + \\
&\quad (0.1)(0.4) + (0.1)(0.8) + (0.1)(0.8) + (0.1)(0.9) \\
&= 0.33
\end{aligned}$$

and likewise we find that $\tilde{p}_2 = 0.67$. Therefore, Procedure AVC is just a specific case of our generalized method whereby each observation is equally likely to occur. In general, the multinomial probabilities found using Procedure AVC can be calculated using our method by

$$\tilde{p}_j = \sum_{i=1}^{v} \left[\left(\frac{1}{v} \right) \prod_{\ell=1;\, \ell \neq j}^{k} \left(\Pr\{X_\ell \leq x_{j(i)}\} \right) \right] \tag{12}$$

for $j = 1, 2, \ldots k$.

However, with perfect knowledge of the underlying population distributions of Policy A and Policy B, $\tilde{p}_1$ and $\tilde{p}_2$ are the exact multinomial probabilities for each policy. Hence, we can see that the estimates using Procedure BEM and Procedure AVC are just that – empirical estimates which approach the true probabilities as determined by the integral method.

Therefore, in general, the multinomial probability that system j will be the best in a single comparison among k systems is computed as follows:

$$\tilde{p}_j = \int_{-\infty}^{\infty} f_j(x) \prod_{\ell=1;\, \ell \neq j}^{k} F_\ell(x)\, dx \tag{13}$$

where $f_j(x)$ is the probability density function for system j, $j = 1, 2, \ldots, k$ and $F_\ell(x)$ is the cumulative distribution function of system ℓ, $\ell = 1, 2, \ldots, k;\ \ell \neq j$.

There are multiple advantages of the integral method compared with its BEM and AVC counterparts. As long as we have knowledge of the underlying theoretical distributions of the parameter of interest from each system, the integral method provides an exact value for the multinomial probabilities. For cases where there is no closed form solution to equation (13), some form of numerical integration could be performed to provide a sufficiently accurate estimate. Our method also requires only one simple calculation for each multinomial probability as opposed to the v comparisons made in Procedure BEM or even worse, the v^k comparisons in Procedure AVC. Furthermore, if we know the underlying distribution of the parameter of interest for each system, we don't need to collect any data from the systems to calculate empirical estimates as done in the other procedures. There is no need to determine the minimum sample size required to achieve a certain level of fidelity in our probability calculations. With the integral method, we simply calculate the multinomial probabilities based on the distributions themselves. In a real world application, this could mean huge savings in both time and money.

There are also some disadvantages inherent with the integral method. This method hinges upon the knowledge of the underlying distributions of the parameter of interest for comparing the competing systems. Often times, especially when testing a new system from which we have collected little or no data, we may have very little, if any, knowledge of the underlying distribution. Unless there is a subject matter expert who knows what the distribution is or has historical data to analyze, we won't be able to use this method in its purest form. However, if this is the case, we don't have to completely abandon the method altogether. As we saw using the example data in Table 1, we can still calculate an empirical estimate of the multinomial probabilities as long as we can estimate the pdf and cdf of each policy. Unfortunately, it is this reality of imperfect knowledge of the underlying theoretical distributions of the parameter of interest that leads us into the next section of this paper where we conduct a study comparing the results of Procedure BEM and Procedure AVC with the empirical probability estimates using the generalized integral method.

2.5 Empirical Comparison Study

In the previous section, we discussed the need for calculating empirical estimates of the multinomial probabilities in real world situations. For the purposes of this study, we compared the empirical probability estimates found using Procedure BEM and Procedure AVC with those of the proposed integral method. The empirical estimates found using each of the three methods were also compared to truth values as determined by the exact multinomial probabilities computed using the generalized integral method in equation (13). Sample data were generated using the Arena Input Analyzer® software package from Rockwell Software for seven pairs of distributions. Empirical estimates of the multinomial probabilities, p_1 and p_2, were calculated using Procedure BEM, Procedure AVC, and the integral method. Table 4 lists the pairs of distributions analyzed, along with their associated indifference zone parameter value, θ, where $\theta = \frac{\tilde{p}_2}{\tilde{p}_1}$.

Table 4. Empirical Comparison Study Test Cases

Test Case	System 1	System 2	θ
1	$\text{Beta}(\alpha = 1, \beta = 3)$	$\text{Beta}(\alpha = 1, \beta = 2.5)$	1.2
2	$\text{Beta}(\alpha = 2, \beta = 5)$	$\text{Beta}(\alpha = 2, \beta = 2)$	3.4
3	$\text{Exponential}(\mu = 1)$	$\text{Exponential}(\mu = 1.2)$	1.2
4	$\text{Exponential}(\mu = 1)$	$\text{Exponential}(\mu = 2.2)$	2.2
5	$\text{Exponential}(\mu = 1)$	$\text{Exponential}(\mu = 3.4)$	3.4
6	$\text{Gamma}(\alpha = 2, \beta = 2)$	$\text{Gamma}(\alpha = 3, \beta = 2)$	2.2
7	$\text{Normal}(\mu = 0, \sigma = 1)$	$\text{Normal}(\mu = 0.5, \sigma = 1)$	1.8

For each of the seven cases, we compared the empirical probability estimates across a range of sample sizes, $v = 10, 30,$ and 100. In each combination of distribution comparison and sample size, 1024 macro-replications were conducted.

In order to calculate empirical estimates for the multinomial probabilities using the integral method, we need to be able to estimate the pdf and cdf of the parameter of interest for each system. For the purpose of this analysis, when calculating the multinomial probability, $\tilde{p}_j$, the empirical cdf for each system is simply calculated as

$$F_\ell(x_i) = \frac{\text{number of elements from system } \ell \ \leq x_i}{v} \tag{14}$$

for each system ℓ, where $\ell \neq j$ and $i = 1, 2, \ldots, v$. For example, in the two sample case when calculating $\tilde{p}_1$,

$$F_2(x_i) = \frac{\text{number of elements from system } 2 \ \leq x_i}{v},$$

for each x_i from system 1.

Now we just need to estimate the pdf for each system. The histogram method, as previously shown at the beginning of section 2.4, is one of the most common density estimation techniques. However, the accuracy of the histogram technique is heavily dependent on the selection of the bin width and the location of the bin end points. The empirical pdf values can vary greatly based on the choice these parameter values. Therefore, instead of using the histogram method in order to estimate the pdf, we have elected to use kernel density estimation to calculate the multinomial probability estimates, $\tilde{p}_1$ and $\tilde{p}_2$, for each of the seven test cases. It is important to note here again, if we use the individual observations to construct our empirical pdfs, the multinomial probabilities found using the generalized integral method match the results using Procedure AVC.

Like the histogram method, kernel density estimation is a nonparametric density estimation technique which is useful for small data sets. The roots of kernel density estimation can be found in two seminal papers by Rosenblatt [19] and Parzen [20], although the basic principles were introduced by Fix and Hodges [21] and Akaike [22]. In general, kernel density estimation is a data smoothing procedure which allows us to move from estimating every density with a step function by making more efficient use of the data in order to construct a smooth estimate of the density function. Figure 7 below shows a typical density estimation for six data points using the histogram as well as a smooth kernel density estimate using the same six data points.

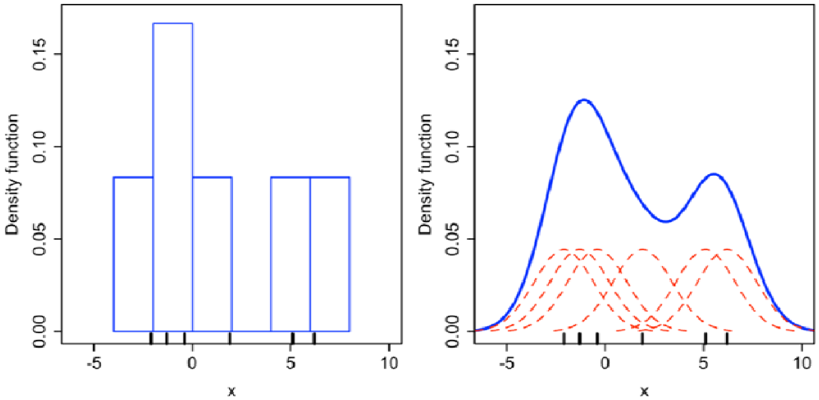

Figure 7. Comparison of the histogram (left) and kernel density estimate (right) using the same data

Instead of binning each data point into a bin of arbitrary size and location, the kernel density estimation technique represents each data point by an individual kernel (in this case a normal distribution) centered at each observation as represented by the dashed curves in Figure 7. These six individual distributions are then accumulated into a smooth kernel density estimate. Therefore, in areas where there are more observations, the kernel density will assume a larger value than in areas where there are relatively few observations.

Formally, the kernel density estimator is defined as

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{15}$$

where $K(\cdot)$ is a function satisfying $\int K(x)dx = 1$, which we call the kernel, and h is a positive number called the bandwidth [23]. The kernel determines the shape of the function and the bandwidth acts as a smoothing parameter. The kernel is usually a unimodal probability density function that is symmetric about zero, which ensures that $\hat{f}(x;h)$ is also a density [23]. Many different kernel functions are used, such as uniform, triangular, Gaussian, cosine, and numerous others functions. For the purpose of our Monte Carlo simulation, we will use the common Gaussian kernel which is defined as

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2} . \tag{16}$$

It should be noted that the choice of the kernel function is not nearly as important as the choice of the value for the bandwidth, and incorrect selection can result in under or over smoothing of the kernel density estimator. When the bandwidth value is too small and the estimate pays too much attention to the data set and doesn't allow for variation across samples, the density curve becomes noisy and is said to be undersmoothed. When the bandwidth value is too large, the natural structure of the underlying density is smoothed away, the estimate is said to be oversmoothed. However, with a correct selection of h, the kernel estimate is not overly noisy and the essential structure of the underlying density is preserved. Figure 8 from Wand and Jones [23] shows an example of each case, where (a) represents an undersmoothed estimate, (b) depicts an oversmoothed estimate, and (c) illustrates a correct bandwidth selection.

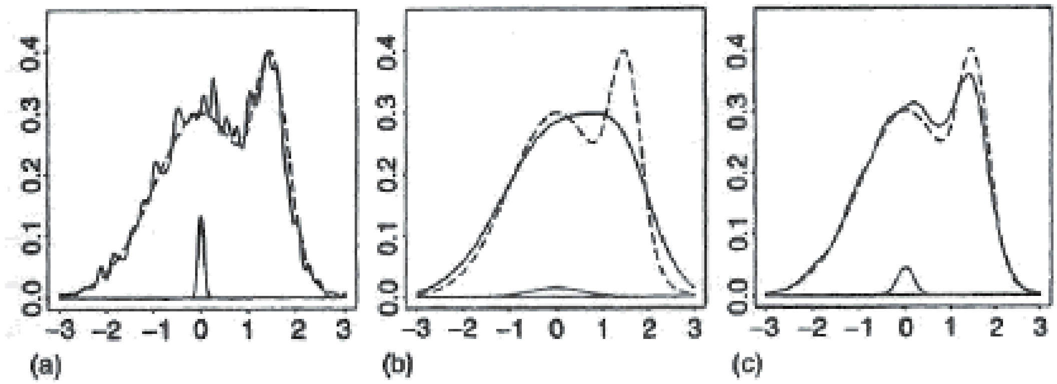

Figure 8. Kernel density estimates for various bandwidths.
The solid line is the density estimate, the dashed line is the true density.
The bandwidths are (a) h = 0.06, (b) h = 0.54, and (c) h = 0.18.

There are numerous approaches used for determining the correct value for bandwidth. If the user has reason to believe that there is certain structure to the data and has knowledge of the position of the modes, it may be acceptable to begin with a large bandwidth and decrease the amount of smoothing until fluctuations in the estimate cease to appear more structural than random [23]. Many other more rigorous tactics exist which make use of automatic bandwidth selectors that produce a bandwidth value based on the sample data itself. For the purpose of our Monte Carlo simulation, we will use the Silverman's rule of thumb [24], which is defined as

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}\,. \tag{17}$$

In order to calculate estimates for the multinomial probabilities, $p_1$and p_2, in our empirical comparison study, we employed built in kernel density estimation functions using the computational software package, *Mathematica 8.0*. We used a Gaussian kernel and bandwidths as determined by Silverman's rule of thumb for all test cases. After generating kernel density estimates for each system, the empirical multinomial probabilities, $\tilde{p}_1$and $\tilde{p}_2$, were then calculated using equation (13).

To establish truth data for comparing the empirical probability estimates found using Procedure BEM, Procedure AVC, and the integral method, we first calculated the exact multinomial probabilities, $\tilde{p}_1$and $\tilde{p}_2$, for each of the seven test cases using the known sample distributions and equation (13). These exact multinomial probabilities (to four decimal places) are shown below in Table 5 along with their associated indifference zone parameter value, $\theta,$ where $\theta = \frac{\tilde{p}_2}{\tilde{p}_1}$.

Table 5. Exact Multinomial Probabilities for Each Test Case

Test Case	$\tilde{p}_1$	$\tilde{p}_2$	θ
1	0.4545	0.5455	1.2
2	0.2262	0.7738	3.4
3	0.4545	0.5455	1.2
4	0.3125	0.6875	2.2
5	0.2273	0.7727	3.4
6	0.3125	0.6875	2.2
7	0.3618	0.6382	1.8

For each of the seven test cases, a summary table was constructed comparing the multinomial probability estimates found using Procedures BEM and AVC with those of the integral method at each of the sample sizes analyzed. The associated error for each of the empirical estimates, denoted $\varepsilon_{\tilde{p}_1}$ and $\varepsilon_{\tilde{p}_2}$, were also calculated as the difference between the probability estimates found using each method and their respective exact multinomial probabilities given in Table 5. The empirical probability estimate summary and associated error summary tables for test cases 1 are presented below in Table 6 and Table 7, respectively.

Table 6. Empirical Probability Estimates for Test Case 1: Beta(α = 1,β = 3) vs. Beta(α = 1,β = 2.5), $\boldsymbol{\theta}$ = 1.2

Method	Sample Size (v)						Average Probabilities	
	10		30		100			
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.4589	0.5411	0.4567	0.5433	0.4530	0.5470	0.4562	0.5438
AVC	0.4586	0.5414	0.4556	0.5444	0.4534	0.5466	0.4559	0.5441
Kernel Density	0.4612	0.5381	0.4589	0.5407	0.4555	0.5440	0.4585	0.5410

Table 7. Empirical Probability Estimate Errors for
Test Case 1: Beta(α = 1,β = 3) vs. Beta(α = 1,β = 2.5), θ = 1.2

Method	Sample Size (v)						Overall Average Error	
	10		30		100			
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0043	0.0043	0.0022	0.0022	0.0016	0.0016	0.0027	0.0027
AVC	0.0041	0.0041	0.0011	0.0011	0.0011	0.0011	0.0021	0.0021
Kernel Density	0.0066	0.0073	0.0043	0.0048	0.0009	0.0014	0.0040	0.0045

We can see from Table 6 that the probability estimates for $\tilde{p}_1$ and $\tilde{p}_2$ are very close between the different methods for each sample size. In fact, the probability estimates are within at most 0.0032 and 0.0037 for $\tilde{p}_1$ and $\tilde{p}_2$, respectively. This same behavior was observed for the other six test cases as well.

Looking at Table 7, the probability estimate errors are fairly small for each method, and in general, the errors decreased as the sample size increased. The errors resulting from Procedure BEM and Procedure AVC are fairly close, with AVC errors consistently smaller. The resulting errors from the kernel density estimation technique, however, were larger than all other techniques. These larger errors, although still quite small, were due to outlying estimates found using the kernel density estimation technique. The remaining probability estimate and error summary tables for the other six test cases can be found in Appendix A.

Table 8 below summarizes the average errors for each method at each of the three sample sizes, across all seven test cases, as well as the overall average errors for each method.

Table 8. Average Empirical Probability Estimate Errors
Across All 7 Test Cases For Each Method Based on Sample Size

Method	Sample Size (v)						Overall Average Error	
	10		30		100			
	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0038	0.0041	0.0024	0.0021	0.0010	0.0010	0.0024	0.0024
AVC	0.0027	0.0027	0.0018	0.0018	0.0010	0.0010	0.0018	0.0018
Kernel Density	0.0164	0.0165	0.0140	0.0142	0.0100	0.0101	0.0135	0.0136

Again, we see that the average errors for Procedure BEM and Procedure AVC are very similar, with AVC errors consistently smaller. As sample size increases, the average errors decrease for each method, with Procedure AVC performing best at $v = 100$ samples. Yet again, we see that the overall average error for the kernel density estimation technique is much larger than the resulting errors of the other techniques. In fact, the overall average error using the density estimation technique, although perhaps still acceptably small in practice, was approximately 7 times as large as the average overall errors of the other methods.

Instead of comparing the average errors for each method based on the various sample sizes, Table 9 shows the average probability estimate errors for each method based on the four indifference zone parameter values tested.

Table 9. Average Empirical Probability Estimate Errors
For Each Method Based on Indifference Zone Parameter Value

Method	Indifference Zone Parameter (θ)							
	1.2		1.8		2.2		3.4	
	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0025	0.0025	0.0025	0.0025	0.0029	0.0029	0.0017	0.0017
AVC	0.0023	0.0023	0.0004	0.0004	0.0017	0.0017	0.0021	0.0021
Kernel Density	0.0035	0.0037	0.0103	0.0103	0.0176	0.0176	0.0210	0.0211

It is interesting to note that changes in the indifference zone parameter don’t exhibit any noticeable effect on the probability estimate errors for each method. The errors for each method remain fairly constant across each of the four indifference zone parameter levels. Although we may have expected the errors to decrease as the indifference zone parameter value increased, no such trend occurred. It appears that the method itself is the most important factor affecting the probability estimate errors, which, on the whole, are quite small considering these are only empirical estimates at relatively small sample sizes.

2.6 Conclusions

We have presented an effective alternate method for comparing competing systems using a generalized integral method. Building on the work of Bechhofer, Elmaghraby, and Morse [1] and Miller, Nelson, and Reilly [2], the generalized integral method provides the ability to calculate the exact probability that a single system is best among competing systems in a single trial. If one possesses knowledge of the distributions of the parameter of interest for each system, the generalized method allows us to avoid calculating computationally intensive empirical probability estimates and provides more accurate multinomial probabilities that could potentially result in significant savings in both time and money when comparing alternate systems. If distributional knowledge of the systems is unknown, we have shown through the Monte Carlo simulation study that the kernel density estimation technique, although lacking in accuracy when compared with Procedures BEM and AVC, is a viable approach for calculating empirical probability estimates using the generalized integral method. We also found that if the individual observations are used for empirical pdf and cdf generation, then the generalized integral method's performance exactly matches that of Procedure AVC. While further analysis could be conducted in order to find more optimal empirical estimation techniques, our method provides an effective alternate approach to the current methods available as well as the traditional techniques which focus solely on comparisons based on long run performance.

III. A Distribution-Free Ordered Comparisons Test:

Determining If One Random Variable is Stochastically Larger Than Another

3.1 Abstract

This paper proposes the use of a distribution-free ordered comparisons procedure to test whether one population is stochastically larger (or smaller) than another. This procedure is suggested as a useful alternative to the well-known Wilcoxon rank sum and Mann-Whitney tests. A Monte Carlo simulation study is conducted, comparing the methods based on simulated power and type I error. All test cases show a marked improvement in performance of the ordered comparisons test versus the Wilcoxon rank sum test, when testing for stochastic dominance. A table of critical values is presented and a large sample approximation is given.

3.2 Introduction

Many nonparametric techniques exist for the two-sample comparison problem, such as the commonly used Wilcoxon rank sum [3] and Mann-Whitney [4] tests. However, in terms of detecting whether one population is stochastically larger (or smaller) than the other, it appears that these test procedures may not make full use of all the information contained in the sample data. This paper presents a new distribution-free method based on ordered comparisons. The technique gleans useful information about the empirical cumulative distribution functions of the sample data by ordering each sample separately (as opposed to a combined ordering as in the Wilcoxon rank sum test). The ordered sets of observations are then compared and a test statistic is calculated.

Essentially, if sample 1 has significantly more smaller values than sample 2, then, in general, the cumulative distribution function of sample 1 will be monotonically greater than the cumulative distribution function of sample 2. This will result in population 2 being stochastically larger than population 1.

Section 3.3 of this paper provides a brief overview of nonparametric statistics and the advantages over their parametric counterparts. A definition of the distribution-free property and stochastic dominance is provided, followed by a brief summary of the Wilcoxon rank sum and Mann-Whitney tests. Section 3.4 introduces the ordered comparisons test procedure, along with a derivation of the associated critical values. A large-sample approximation is also formulated, followed by a numerical example demonstrating the ordered comparisons technique. Section 3.5 compares the ordered comparisons test to the Wilcoxon rank sum test using a Monte Carlo simulation study; simulated power and type I error results are presented.

3.3 Background

Suppose we have $X_1, \dots, X_m$ random observations from population 1 and $Y_1, \dots, Y_n$ random observations from population 2. When comparing these two populations, say based on their respective means, it is common to use a two-sample t-test (small sample size) or z-test (larger sample size). However, both of these methods require us to assume that the observations are normally distributed; as we know, this is not always a valid assumption and can bring in to question whether we are using the appropriate statistical procedure. Both of these well-known techniques are considered parametric methods in that they apply to problems where the distributions from which the samples are drawn are specified, except for the values of a finite number of parameters (i.e. μ, σ^2, etc.) [25].

Alternatively, suppose we wish to test the hypothesis that the two population distributions are identical, but of unspecified form. In this case, we refer to an area of mathematics known as nonparametric statistics. Nonparametric statistical methods require only general assumptions about the nature of the probability distributions [25]. Nonparametric methods, although perhaps not as widely used or familiar as their parametric counterparts, are very useful techniques and offer numerous advantages over parametric methods.

Hollander and Wolfe [26] discuss various advantages that nonparametric methods enjoy to include:

1. Nonparametric methods require few assumptions about the underlying populations from which the data are obtained. In particular, nonparametric procedures forgo the traditional assumption that the underlying populations are normal.

2. Nonparametric procedures are often quite easy to understand.

3. Nonparametric techniques are often easier to apply than their normal theory counterparts.

4. Nonparametric procedures enable the user to obtain exact p-values for tests and the development of computer software has facilitated fast computation of these values.

5. Nonparametric procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal, and they can potentially be much more efficient than these competitors when the underlying populations are not normal.

6. Nonparametric methods are relatively insensitive to outlying observations.

7. Nonparametric procedures are applicable in many situations where normal theory procedures cannot be utilized.

The distribution-free property is another important aspect associated with many nonparametric procedures. A distribution-free test statistic is used in the Wilcoxon rank sum test, the equivalent Mann-Whitney test, and in the ordered comparisons test proposed in this paper.

Consider the two-sample case where we have an independent random sample of m observations from population 1 with continuous cumulative distribution function, $F_1(x)$, and an independent random sample of n observations from population 2 with continuous cumulative distribution function, $F_2(x)$. The null hypothesis being tested is:

$$H_0\text{: } F_1(x) = F_2(x) = F(x),\ F(x) \text{ unspecified.}$$

We are testing that the two random samples have the same probability distribution, but the common distribution is not specified [26]. In the case of the Wilcoxon rank sum test, the test statistic, W, is simply the sum of the ranks obtained for population 2 in the joint ranking.

When H_0 is true, the distribution of W does not depend on $F(x)$. Hence, when H_0 is true, for all x-values, the probability that $W \leq x$, denoted $P_0(W \leq x)$, does not depend on $F(x)$. This is known as the distribution-free property [26]. This allows us to table the distribution of W under H_0 without specifying or even knowing the underlying probability distribution function, $F(x)$.

When considering the alternative hypothesis for tests such as the Wilcoxon rank sum test or the Mann-Whitney test, we are testing that Y tends to be larger (or smaller) than X. That is, that Y is stochastically larger (or smaller) than X. By definition, the random variable Y with cumulative distribution function $F_2(x)$ is said to be stochastically larger than the random variable X with cumulative distribution function $F_1(x)$ if:

$$F_2(x) \leq F_1(x), \tag{18}$$

for every x, with strict inequality for at least one value [27]. If Y is stochastically larger than X, then X is said to be stochastically smaller than Y and the two distributions are said to be stochastically ordered with X less than Y [27]. For simplification purposes, this is sometimes written as $F_2(\cdot)$ is stochastically larger than $F_1(\cdot)$.

3.3.1 Wilcoxon Rank Sum Test

For the Wilcoxon rank sum test, begin by collecting $N = m + n$ random observations $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, which are independent and identically distributed. We also assume that the X's and Y's are mutually independent and that populations 1 and 2 are continuous populations [26]. Let $F_1(x)$ be the cumulative distribution function for population 1 and $F_2(x)$ be the cumulative distribution function for population 2.

The null hypothesis is:

$$H_0\colon F_1(x) = F_2(x), \text{ for every } x.$$

The Wilcoxon rank sum test statistic, W, is computed by first ordering the combined sample of X-values and Y-values from least to greatest and assigning a rank

$1, \dots, N$ to each ordered term [3]. W is then calculated as the sum of the ranks assigned to the Y-values [26].

Specifically,

$$W = \sum_{i=1}^{n} S_i \tag{19}$$

where S_i denotes the rank of Y_i in the combined ordering [26].

The alternative hypothesis can take one of three forms. Namely,

a.

$$H_1 \text{: } F_2(\cdot) \text{ is stochastically larger than } F_1(\cdot)$$

We reject H_0 at the α level of significance if:

$$W \geq w_\alpha$$

where w_α is chosen to make the type I error probability, P(rejecting H_0 when H_0 is true), equal to α; otherwise, do not reject H_0 [26]. It is important to note that when naming the samples, the Y-sample is the sample with the smaller sample size. If $m = n$, then either sample can be designated as the Y-sample.

b.

$$H_1 \text{: } F_1(\cdot) \text{ is stochastically larger than } F_2(\cdot)$$

We reject H_0 at the α level of significance if:

$$W \leq n(m+n+1) - w_{\alpha}.$$

Otherwise, do not reject H_0.

c.

$$H_1: F_1(x) \neq F_2(x)$$

We reject H_0 at the α level of significance if:

$$W \geq w_{\alpha_2} \text{ or } W \leq n(m+n+1) - w_{\alpha_1}$$

where $\alpha_1 + \alpha_2 = \alpha$. Otherwise, do not reject H_0 [26]. For a table of critical values, w_{α}, see Hollander and Wolfe [26].

3.3.2 Large-Sample Approximation

For sample sizes where $m, n > 10$, we can use a large-sample approximation based on the asymptotic normality of W, standardized. This approximation requires knowledge of the mean and variance of W when the null hypothesis is true. When H_0 is true, the mean and variance of W are, respectively,

$$E_0(W) = \frac{n(m+n+1)}{2} \tag{20}$$

$$\mathrm{var}_0(W) = \frac{mn(m+n+1)}{12} \tag{21}$$

The standardized large-sample approximation of W is:

$$W^* = \frac{W - E_0(W)}{\sqrt{\mathrm{var}_0(W)}} = \frac{W - [n(m+n+1)/2]}{\sqrt{mn(m+n+1)/12}} \tag{22}$$

where W^* has an asymptotic $N(0,1)$ distribution when H_0 is true [26].

The normal theory approximation for alternative hypothesis (a) is:

Reject H_0 if $W^* \geq z_\alpha$; otherwise do not reject.

The normal theory approximation for alternative hypothesis (b) is:

Reject H_0 if $W^* \leq -z_\alpha$; otherwise do not reject.

The normal theory approximation for alternative hypothesis (c) is:

Reject H_0 if $|W^*| \geq z_{\alpha/2}$; otherwise do not reject.

The critical value, z_α, comes from the standard normal distribution.

3.3.3 Mann-Whitney Test

In 1947, H. B. Mann and D. R. Whitney proposed a test similar to Wilcoxon's rank sum test, again testing if one random variable is stochastically larger (or smaller) than the other. It can be shown that the Wilcoxon rank sum test and the Mann-Whitney test produce equivalent results [26]. Mann and Whitney showed that, in the case of no ties, their U-statistic is directly related to Wilcoxon's W-statistic by:

$$U = mn + \frac{m(m+1)}{2} - W \tag{23}$$

where we have m observations from population 1 and n observations from population 2 and W is the rank sum of the associated x-values from population 1 [4].

As in the Wilcoxon rank sum test, we begin by ordering the $N = m + n$ samples from least to greatest. However, instead of summing the ranks of the y-values as in the Wilcoxon rank sum test, we calculate U by counting the number of times a y precedes an x [4]. We then use this U-statistic to test the null hypothesis:

$$H_0\colon F_1(x) = F_2(x), \text{ for every } x$$

just as we did in the Wilcoxon rank sum test. We can also test against the same three alternative hypotheses as the Wilcoxon rank sum test. For a list of the associated rejection criterion for each alternative hypothesis, and a table of associated critical values, and a large-sample approximation formula, see Wackerly, Mendenhall, and Sheaffer [25].

For the remainder of this paper, however, we will make use of the Wilcoxon rank sum test for purposes of comparison with the proposed distribution-free ordered comparisons test.

3.4 Ordered Comparisons Test

We propose a new nonparametric hypothesis test similar to the Wilcoxon rank sum and Mann-Whitney tests for determining if one population is stochastically larger (or smaller than) the other. Let $X_1, \dots, X_m$ be random observations from population 1 and $Y_1, \dots, Y_n$ be random observations from population 2. That is, the X's are independent and identically distributed and the $Y's$ are independent and identically distributed. In addition to the assumptions of independence within each sample, we also assume independence between the two samples. That is, the X's and Y's are mutually independent. Lastly, we will assume that populations 1 and 2 are both continuous populations.

Let $F_1(x)$ be the continuous cumulative distribution function for population 1 and $F_2(x)$ be the continuous cumulative distribution function for population 2. The null hypothesis is:

$$H_0\colon F_1(x) = F_2(x), \text{ for every } x.$$

That is, X and Y have the same probability distribution, but the common distribution is not specified.

3.4.1 Procedure

If $m = n$, begin by separately ordering the observations from each sample from least to greatest. We now have $X_{[1]}, \dots, X_{[m]}$ ordered X-values and $Y_{[1]}, \dots, Y_{[n]}$ ordered Y-values. Let M be the common sample size when $m = n$ or the $\min(m, n)$ when $m \neq n$. [Note: If $m > n$, randomly select n observations from the X-sample data so that we have an equal number of samples from each population. Likewise, if $m < n$, randomly select m observations from the Y-sample data.]

To compute our test statistic L, we compare the ordered pairs of observations and count the number of times that $X_{[i]} < Y_{[i]}$. Specifically,

$$L = \sum_{i=1}^{\mathrm{M}} \varphi\left(X_{[i]}, Y_{[i]}\right) \tag{24}$$

where

$$\varphi\left(X_{[i]}, Y_{[i]}\right) = \begin{cases} 1, & \text{if } X_{[i]} < Y_{[i]} \\ 0, & \text{otherwise} \end{cases},$$

and from our assumptions the probability of a tie is zero. That is, $X_{[i]} \neq Y_{[i]}$, for any i. However, if we don’t make this assumption, then for every ordered replication where a tie exists, randomly select one of the tied systems as best in that replication.

To test

$$H_0\colon F_1(x) = F_2(x), \text{ for every } x,$$

versus

$$H_1\colon F_2(\cdot) \text{ is stochastically larger than } F_1(\cdot)$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } L \geq \ell_\alpha; \quad \text{otherwise do not reject,}$$

where the constant ℓ_α is chosen to make the type I error probability equal to α. Values of ℓ_α based on the sample size, M, are given in Table 43 in Appendix B.

To test

$$H_0\colon F_1(x) = F_2(x), \text{ for every } x,$$

versus

$$H_1\colon F_1(\cdot) \text{ is stochastically larger than } F_2(\cdot)$$

at the α level of significance,

Reject H_0 if $L \leq \mathrm{M} - \ell_\alpha$;   otherwise do not reject.

3.4.2 Derivation of the Distribution of *L*

Assuming that the underlying, unknown distribution under H_0 is continuous, ties among $(X_{[i]}, Y_{[i]})$ have a zero probability of occurring. That is, $X_{[i]} \neq Y_{[i]}$, for any i. Then under H_0, all possible arrangements of the ordered comparisons are equally likely, with probability

$$\frac{1}{\sum_{i=0}^{M} \binom{M}{i}}.$$

For example, consider the simple case where $m = n = M = 4$. There are $\sum_{i=0}^{4} \binom{4}{i} = 16$ possible arrangements of the ordered comparisons, each with a $\frac{1}{16} = .0625$ probability of occurring. Specifically, there is $\binom{4}{0} = 1$ way that $L = 0$, $\binom{4}{1} = 4$ ways that $L = 1$, $\binom{4}{2} = 6$ ways that $L = 2$, $\binom{4}{3} = 4$ ways that $L = 3$, and $\binom{4}{4} = 1$ way that $L = 4$, resulting in 16 possible arrangements. Therefore, the probability that $L = 3$ under H_0, is $\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{4}{16}$. The other probability values are calculated in a similar fashion and shown in Table 10.

Table 10. Null Distribution of L for $M = 4$

Possible Value of L	Probability of Value
0	1/16 = .0625
1	4/16 = .25
2	6/16 = .375
3	4/16 = .25
4	1/16 = .0625

Hence, under H_0, the probability that L is greater than or equal to 3 is:

$$P_0(L \geq 3) = P_0(L = 3) + P_0(L = 4)$$

$$= .25 + .0625 = .3125.$$

Looking at the critical values in Table 43 in Appendix B, we see that this matches the tail probability corresponding to $x = 3$ when $M = 4$. Therefore, if we were to obtain an L-statistic of 3 when $M = 4$, we could reject H_0 at only the $\alpha = .3125$ level of significance.

As with the Wilcoxon rank sum and Mann-Whitney tests, the distribution derived for our test statistic does not depend on the underlying distributions of the two populations. This makes our test a distribution-free procedure as previously discussed.

3.4.3 Large-Sample Approximation

In order to construct the large-sample approximation of L, we must first determine the mean and variance of L under the null hypothesis. The mean of L under H_0 is:

$$E_0[L] = E_0\left[\sum_{i=1}^{M} \varphi\left(X_{[i]}, Y_{[i]}\right)\right] \tag{25}$$

and because the expectation of the sum of independent trials is equal to the sum of the expectations [25],

$$E_0\left[\sum_{i=1}^{M} \varphi\left(X_{[i]}, Y_{[i]}\right)\right] = \sum_{i=1}^{M} E_0\left[\varphi\left(X_{[i]}, Y_{[i]}\right)\right]$$

and by the definition of expectation and $\varphi\left(X_{[i]}, Y_{[i]}\right)$ we have

$$\sum_{i=1}^{M} E_0\left[\varphi\left(X_{[i]}, Y_{[i]}\right)\right] = \sum_{i=1}^{M}[1 \cdot \mathrm{Prob}(\varphi = 1) + 0 \cdot \mathrm{Prob}(\varphi = 0)]$$

$$= \sum_{i=1}^{M}[1 \cdot \mathrm{Prob}(\varphi = 1)]$$

and since under the null hypothesis $F_1(x) = F_2(x)$, for every x, we know that

$$\sum_{i=1}^{M}[1 \cdot \mathrm{Prob}(\varphi = 1)] = \sum_{i=1}^{M}\left[\frac{1}{2}\right] = \frac{M}{2}$$

Therefore, we have shown that

$$E_0[L] = \frac{M}{2}. \tag{26}$$

The variance of L is:

$$\operatorname{var}_0[L] = \operatorname{var}_0\left[\sum_{i=1}^{M} \varphi\left(X_{[i]}, Y_{[i]}\right)\right] \tag{27}$$

and using the fact that all the φ's are independent,

$$\operatorname{var}_0\left[\sum_{i=1}^{M} \varphi\left(X_{[i]}, Y_{[i]}\right)\right] = \sum_{i=1}^{M} \operatorname{var}_0\left[\varphi\left(X_{[i]}, Y_{[i]}\right)\right]$$

and since for any of the φ's, $\varphi^2 = \varphi$,

$$E_0\left[\varphi^2\left(X_{[i]}, Y_{[i]}\right)\right] = E_0\left[\varphi\left(X_{[i]}, Y_{[i]}\right)\right]$$

and from the previous derivation of $E_0[L]$ we know that

$$E_0\left[\varphi\left(X_{[i]}, Y_{[i]}\right)\right] = \frac{1}{2}$$

and thus

$$\begin{aligned}\operatorname{var}_0[\varphi(X_{[i]},Y_{[i]})] &= E_0[\varphi^2(X_{[i]},Y_{[i]})] - \{E_0[\varphi(X_{[i]},Y_{[i]})]\}^2 \\ &= E_0[\varphi(X_{[i]},Y_{[i]})] - \{E_0[\varphi(X_{[i]},Y_{[i]})]\}^2 \\ &= \frac{1}{2} - \left\{\frac{1}{2}\right\}^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}\end{aligned}$$

and by substitution we have

$$\sum_{i=1}^{M} \operatorname{var}_0[\varphi(X_{[i]},Y_{[i]})] = \sum_{i=1}^{M} \frac{1}{4} = \frac{M}{4}$$

Therefore, we have shown that

$$\operatorname{var}_0[L] = \tfrac{M}{4}. \tag{28}$$

The standardized large-sample approximation of L, based on the asymptotic normality of L is then defined as

$$L^* = \frac{L - E_0[L]}{\sqrt{\operatorname{var}_0[L]}} = \frac{L - \frac{M}{2}}{\sqrt{\frac{M}{4}}} = \frac{L - \frac{M}{2}}{\frac{\sqrt{M}}{2}} = \frac{2L - M}{\sqrt{M}} \tag{29}$$

and the central limit theorem establishes that as $M \to \infty$, L^* has a limiting $N(0,1)$ distribution [28].

To test

$$H_0: F_1(x) = F_2(x), \text{ for every } x,$$

versus

$$H_1: F_2(\cdot) \text{ is stochastically larger than } F_1(\cdot)$$

at the α level of significance, the normal theory approximation is:

$$\text{Reject } H_0 \text{ if } L^* \geq z_\alpha; \quad \text{otherwise do not reject.}$$

To test

$$H_0: F_1(x) = F_2(x), \text{ for every } x,$$

versus

$$H_1: F_1(\cdot) \text{ is stochastically larger than } F_2(\cdot)$$

at the α level of significance, the normal theory approximation is:

$$\text{Reject } H_0 \text{ if } L^* \leq -\mathrm{z}_\alpha; \quad \text{otherwise do not reject.}$$

3.4.4 Example Calculation

Let's now provide a simple example, illustrating the ordered comparisons test procedure. Consider again the following random sample data used in Section 2.3.3 where $m = n = M = 10$:

Table 11. Example Data

X-values	Y-Values
0.31950	0.48383
0.76029	1.18579
0.26898	0.83970
1.40037	0.87125
3.49909	0.47615
0.01253	2.19601
0.27408	0.62893
0.10418	4.42322
1.55529	0.52177
0.81182	1.03523

Begin by separately ordering the observations from each sample from least to greatest. Then calculate $\varphi(X_{[i]}, Y_{[i]})$ for each ordered pair of observations. That is,

$$\varphi(X_{[i]}, Y_{[i]}) = \begin{cases} 1, & \text{if } X_{[i]} < Y_{[i]} \\ 0, & \text{otherwise} \end{cases}$$

and

$$L = \sum_{i=1}^{10} \varphi(X_{[i]}, Y_{[i]})$$

Table 12. Example L-Statistic Calculation

X-values	Y-Values	$\varphi\left(X_{[i]}, Y_{[i]}\right)$
0.01253	0.47615	1
0.10418	0.48383	1
0.26898	0.52177	1
0.27408	0.62893	1
0.31950	0.83970	1
0.76029	0.87125	1
0.81182	1.03523	1
1.40037	1.18579	0
1.55529	2.19601	1
3.49909	4.42322	1
		L = 9

We will test

$$H_0: F_1(x) = F_2(x), \text{ for every } x,$$

versus

$$H_1: F_2(\cdot) \text{ is stochastically larger than } F_1(\cdot)$$

at the $\alpha \approx$ .05 level of significance where we

$$\text{Reject } H_0 \text{ if } L \geq \ell_\alpha; \quad \text{otherwise do not reject.}$$

Looking at the critical values in Table 43 in Appendix B, we find that for $m = n = M = 10$, $\ell_\alpha = 8$ at the $\alpha =$ .0547 level of significance. Therefore, since

$(L = 9) > (\ell_{.0547} = 8)$, we can reject H_0 at the $\alpha = .0547$ level of significance and conclude that there is significant statistical evidence to suggest that $F_2(\cdot)$ is stochastically larger than $F_1(\cdot)$.

3.4.5 The Probability That One System Is Stochastically Larger Than Another

It is worth noting that we can use the test statistic L calculated during the ordered comparisons test in order to estimate the probability that system 2 is stochastically larger than system 1. Let ρ be the probability that system 2 is stochastically larger than system 1. That is, let $\rho = \Pr\{F_2(x) \leq F_1(x)\}$ for all x, with strict inequality at some x, or in notation form, let ρ be the probability that $\Pr\{\text{System } 2 \geq x\} \geq \Pr\{\text{System } 1 \geq x\}$ for all x, and for some x, $\Pr\{\text{System } 2 \geq x\} > \Pr\{\text{System } 1 \geq x\}$. Using the L statistic from the ordered comparisons test and the common sample size, M, we can compute a point estimator for ρ, given by

$$\hat{\rho} = {}^{L}\!/_{M}. \tag{30}$$

It can be shown that

$$E[\hat{\rho}] = E\left[\frac{L}{M}\right] = \frac{E[L]}{E[M]} = \frac{{}^{M}\!/_{2}}{M} = \frac{1}{2} = \rho. \tag{31}$$

Therefore, the point estimator $\hat{\rho}$ is an unbiased estimator for the probability that system 2 is stochastically larger than system 1. On the other hand, $1 - \hat{\rho}$ is an unbiased estimator for the probability that system 1 is stochastically larger than system 2.

Looking at the L statistic calculated in Table 12 for the preventative maintenance policy data, we can calculate the unbiased estimate for the probability that Policy B is stochastically larger than Policy A as

$$\hat{\rho} = {}^{L}\!/_{M} = {}^{9}\!/_{10} = 0.9.$$

3.5 Empirical Comparison Study

In order to characterize the performance of the ordered comparisons test, a Monte Carlo simulation study was conducted. The ordered comparisons test was compared to the Wilcoxon rank sum test based on the simulated power (the proportion of correct rejections of H_0) and type I error rates. Seven two-sample cases were analyzed, where the second population is known to be stochastically larger than the first. Therefore, $\tilde{p}_2 > \tilde{p}_1$ where $\tilde{p}_j$ is defined in Section 2.4, equation (13) as:

$$\tilde{p}_j = \int_{-\infty}^{\infty} f_j(x) \prod_{\ell=1;\, \ell \neq j}^{k} F_\ell(x)\, dx$$

where $f_j(x)$ is the probability density function for system j, $j = 1, 2, \dots, k$ and $F_\ell(x)$ is the cumulative distribution function of system ℓ, $\ell = 1, 2, \dots, k;\ \ell \neq j$.

As a measure of how much better the second system is than the first, we define the indifference zone parameter as $\theta = \frac{\tilde{p}_2}{\tilde{p}_1}$. The following table lists the pairs of distributions analyzed along with their associated θ value.

Table 13. Empirical Comparison Study Test Cases

Test Case	System 1	System 2	θ
1	Beta(α = 1, β = 3)	Beta(α = 1, β = 2.5)	1.2
2	Beta(α = 2, β = 5)	Beta(α = 2, β = 2)	3.4
3	Exponential(μ = 1)	Exponential(μ = 1.2)	1.2
4	Exponential(μ = 1)	Exponential(μ = 2.2)	2.2
5	Exponential(μ = 1)	Exponential(μ = 3.4)	3.4
6	Gamma(α = 2, β = 2)	Gamma(α = 3, β = 2)	2.2
7	Normal(μ = 0, σ = 1)	Normal(μ = 0.5, σ = 1)	1.8

For each of the seven cases, we examined test performance of the ordered comparisons test versus the Wilcoxon rank sum test across a range of sample sizes, $M = 10, 30, 50,$ and 100. The sample data were generated using the Arena Input Analyzer® software package from Rockwell Software. In each combination of distribution comparison and sample size, 1024 macro-replications were produced in order to reduce the Monte Carlo error.

Using the ordered comparisons test, we tested

$$H_0\colon F_1(x) = F_2(x), \text{ for every } x,$$

versus

$$H_1\text{: } F_2(\cdot) \text{ is stochastically larger than } F_1(\cdot)$$

at the $\alpha \approx .05$ level of significance, where we

$$\text{Reject } H_0 \text{ if } L \geq \ell_{\alpha\approx.05}; \quad \text{otherwise do not reject.}$$

Using the ordered comparisons critical values from Table 43 in Appendix B, we see that the critical values for the various sample sizes are

$$\begin{aligned} \ell_{\alpha=.0547} &= 8, \quad m = n = M = 10 \\ \ell_{\alpha=.0494} &= 20, \quad m = n = M = 30 \\ \ell_{\alpha=.0595} &= 31, \quad m = n = M = 50 \\ \ell_{\alpha=.0443} &= 59, \quad m = n = M = 100 \end{aligned}$$

For the Wilcoxon rank sum test, we tested the same null and alternate hypotheses, but we rejected H_0 at the $\alpha \approx .05$ level of significance if $W \geq w_\alpha$. For the case where $M = 10$, $w_{\alpha=.0446} = 128$. For the other three sample sizes, we used the large-sample approximation:

$$W^* = \frac{W - E_0(W)}{\sqrt{Var_0(W)}} = \frac{W - [n(m+n+1)/2]}{\sqrt{mn(m+n+1)/12}}$$

where W^* has an asymptotic $N(0,1)$ distribution when H_0 is true, and we rejected H_0 if:

$$W^* \geq z_{.05}; \quad \text{otherwise we did not reject.}$$

The power results are presented below:

Table 14. Simulated Power ($\alpha \approx .05$) for
Test Case 1: Beta($\alpha = 1, \beta = 3$) vs. Beta($\alpha = 1, \beta = 2.5$), $\theta = 1.2$

Sample Size	Power		
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons	
10	0.106	0.403	
30	0.158	0.579	0.579*
50	0.205	0.660	0.660*
100	0.322	0.807	0.807*

*using the large-sample approximation for L^*

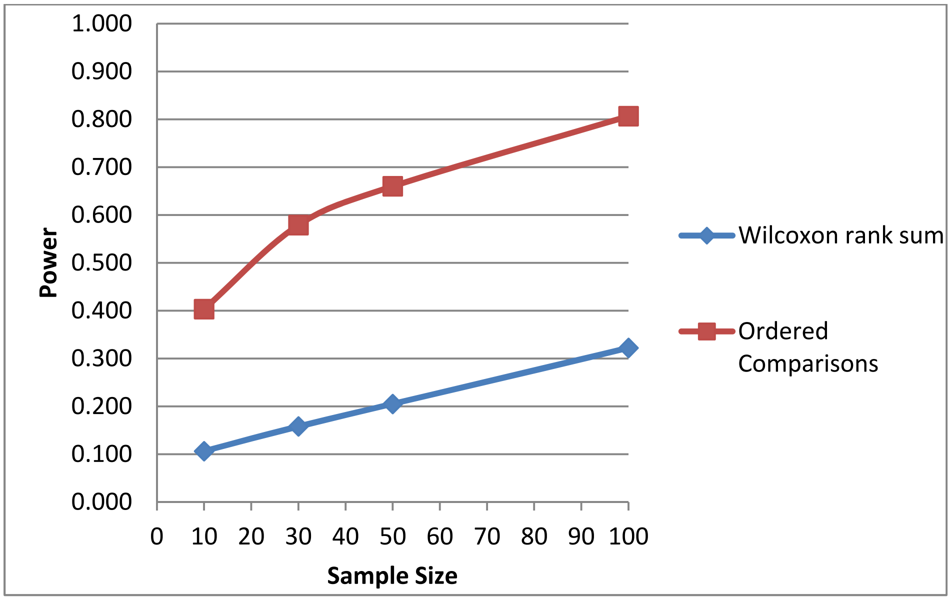

Figure 9. Simulated Power ($\alpha \approx .05$) for
Test Case 1: Beta($\alpha = 1, \beta = 3$) vs. Beta($\alpha = 1, \beta = 2.5$), $\theta = 1.2$

Note that the performance results were identical when using the large-sample approximation versus the critical values found in Table 43 for the ordered comparisons test. This was true for all other cases as well. Therefore, the remaining results are presented using only the ℓ_α critical values as previously discussed. Figures for test cases 2-7 can be found in Appendix C.

Table 15. Simulated Power ($\alpha \approx .05$) for
Test Case 2: Beta($\alpha = 2, \beta = 5$) vs. Beta($\alpha = 2, \beta = 2$), θ = 3.4

Sample Size	Power	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.714	0.929
30	0.993	1.000
50	1.000	1.000
100	1.000	1.000

Table 16. Simulated Power ($\alpha \approx .05$) for
Test Case 3: Exponential ($\mu = 1$) vs. Exponential ($\mu = 1.2$), θ = 1.2

Sample Size	Power	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.108	0.393
30	0.176	0.596
50	0.221	0.694
100	0.328	0.813

Notice in test cases 1 and 3 where $\theta = 1.2$, the power results are nearly identical across all sample sizes even though case 1 is comparing beta distributions and case 3 is comparing exponential distributions. The average difference in power between the two cases across all sample sizes is only 0.010 for the Wilcoxon rank sum test and 0.012 for the ordered comparisons test. It is also worth noting that the ordered comparisons test has a 204% average increase in power over the Wilcoxon rank sum test for cases 1 and 3.

Table 17. Simulated Power ($\alpha \approx .05$) for Test Case 4: Exponential ($\mu = 1$) vs. Exponential ($\mu = 2.2$), $\theta = 2.2$

Sample Size	Power	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.400	0.782
30	0.813	0.979
50	0.949	0.996
100	0.999	1.000

Table 18. Simulated Power ($\alpha \approx .05$) for Test Case 5: Exponential ($\mu = 1$) vs. Exponential ($\mu = 3.4$), $\theta = 3.4$

Sample Size	Power	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.695	0.934
30	0.988	0.999
50	1.000	1.000
100	1.000	1.000

Looking at the three exponential cases, we see that as the difference between the two systems being compared increases (as θ increases), both methods achieve greater power. However, in the third test case where $\theta = 1.2$, the ordered comparisons test has an average increase in power of .416 which amounts to a 200% average increase in power over the Wilcoxon rank sum test. Likewise, the statistical power of both tests also increases as sample size increases.

Table 19. Simulated Power ($\alpha \approx .05$) for
Test Case 6: Gamma ($\alpha = 2, \beta = 2$) vs. Gamma ($\alpha = 3, \beta = 2$), θ = 2.2

Sample Size	Power	
(m = n = M)	Wilcoxon rank sum	Ordered Comparisons
10	0.421	0.799
30	0.817	0.918
50	0.951	0.998
100	0.999	1.000

It is also interesting to compare the results for test cases 4 and 6, where $\theta = 2.2$ in both cases. The power results for both statistical tests are almost identical across all sample sizes even though case 4 is comparing two exponential distributions and case 6 is comparing two gamma distributions. The average difference in power between the two cases across all sample sizes is only 0.007 for the Wilcoxon rank sum test and 0.011 for the ordered comparisons test.

Table 20. Simulated Power ($\alpha \approx .05$) for
Test Case 7: Normal ($\mu = 0, \sigma = 1$) vs. Normal ($\mu = 0.5, \sigma = 1$), θ = 1.8

Sample Size	Power	
(m = n = M)	Wilcoxon rank sum	Ordered Comparisons
10	0.252	0.661
30	0.588	0.934
50	0.771	0.985
100	0.966	1.000

As we can see, the ordered comparisons test has greater power than the Wilcoxon rank sum test for all samples sizes and across all seven test cases. Most notably, the ordered comparison test has a much larger simulated power at the smaller sample sizes. For example, with a sample size of M = 10, the ordered comparisons test has an 81.68%

increase in power over the Wilcoxon rank sum test. Table 21 below summarizes the average increase in power for the ordered comparisons test for all sample sizes analyzed.

Table 21. Average Simulated Power Increase for the Ordered Comparisons Test over the Wilcoxon Rank Sum Test Based on Sample Size

Sample Size ($m = n = M$)	Average Power		Average Increase in Power	Average Percent Increase in Power
	Wilcoxon Rank Sum Test	Ordered Comparisons Test		
10	0.385	0.700	0.315	81.68%
30	0.648	0.858	0.210	32.44%
50	0.728	0.905	0.177	24.25%
100	0.802	0.946	0.144	17.92%

Furthermore, the ordered comparisons test had had an average power of 0.852 over all test cases and sample sizes, whereas the Wilcoxon rank sum test had an average power of 0.641. Therefore, in our Monte Carlo simulation study, we can say that, on average, the ordered comparisons test was 33% more powerful than the Wilcoxon rank sum test and had an average increase in power of 0.211.

Likewise, if we compare the two methods based on the indifference zone parameter, θ, we find that as the difference between the two systems being compared decreases (as θ increases) and it's harder to tell a difference between the systems, the ordered comparisons test performs increasingly better than the Wilcoxon rank sum test. Therefore, in cases where there are only small differences between systems yet we would still like to determine which system is best, the ordered comparison test performs much better than the Wilcoxon rank sum test. Table 22 summarizes the average increase in power of the ordered comparisons test for all four indifference zone parameter levels analyzed.

Table 22. Average Simulated Power Increase for the Ordered Comparisons Test over the Wilcoxon Rank Sum Test Based on the Indifference Zone Parameter ($\boldsymbol{\theta}$)

Indifference Zone Parameter (θ)	Average Power		Average Increase in Power	Average Percent Increase in Power
	Wilcoxon Rank Sum Test	Ordered Comparisons Test		
1.2	0.203	0.618	0.415	204.3%
1.8	0.644	0.895	0.251	38.92%
2.2	0.794	0.934	0.140	17.67%
3.4	0.924	0.983	0.059	6.369%

The simulated type I error rates were determined by analyzing data randomly sampled from the same population for the X-sample and Y-sample. The flowing population distributions were used:

$$\text{Beta}(\alpha = 1, \beta = 3)$$

$$\text{Exponential}(\mu = 2)$$

$$\text{Gamma}(\alpha = 2, \beta = 2)$$

$$\text{Normal}(\mu = 0, \sigma = 1).$$

Since we know that the null hypothesis

$$H_0: F_1(x) = F_2(x), \text{ for every } x$$

is true, the simulated type I error rate is simply the proportion of rejections of H_0 found during the 1024 macro-replications. The simulated type I error rate results are presented below:

Table 23. Simulated Type I Error ($\alpha \approx .05$) for Beta ($\alpha = 1, \beta = 3$) vs. Beta ($\alpha = 1, \beta = 3$)

Sample Size	Type I Error	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.082	0.251
30	0.094	0.341
50	0.100	0.398
100	0.109	0.404

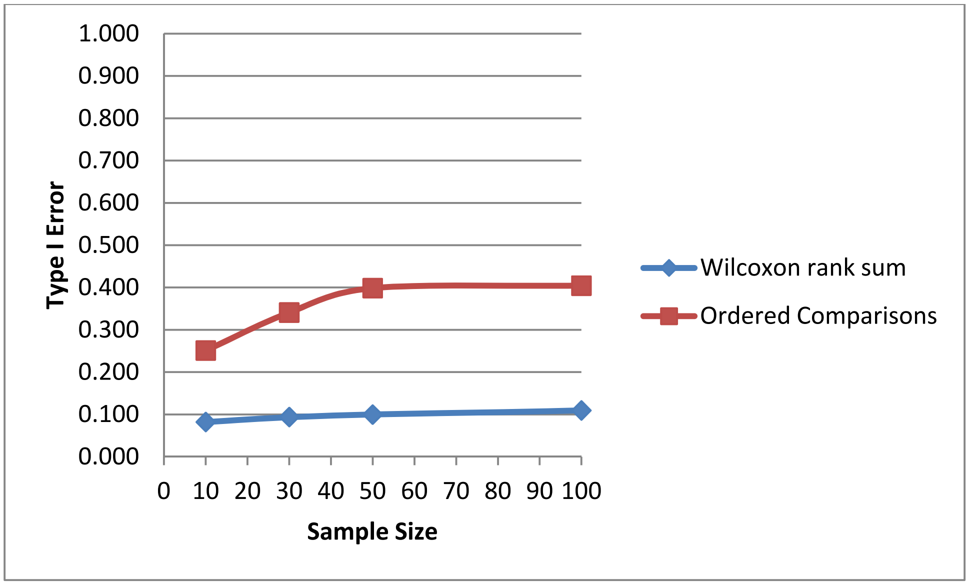

Figure 10. Simulated Type I Error for ($\alpha \approx .05$) for Beta ($\alpha = 1, \beta = 3$) vs. Beta ($\alpha = 1, \beta = 3$)

The Figures for the remaining cases can be found in Appendix C.

Table 24. Simulated Type I Error ($\alpha \approx .05$) for Exponential ($\mu = 2$) vs. Exponential ($\mu = 2$)

Sample Size	Type I Error	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.100	0.268
30	0.105	0.349
50	0.101	0.391
100	0.102	0.401

Table 25. Simulated Type I Error ($\alpha \approx .05$) for Gamma ($\alpha = 2, \beta = 2$) vs. Gamma ($\alpha = 2, \beta = 2$)

Sample Size	Type I Error	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.098	0.300
30	0.098	0.374
50	0.090	0.382
100	0.127	0.426

Table 26. Simulated Type I Error ($\alpha \approx .05$) for Normal ($\mu = 0, \sigma = 1$) vs. Normal ($\mu = 0, \sigma = 1$)

Sample Size	Type I Error	
($m = n = M$)	Wilcoxon rank sum	Ordered Comparisons
10	0.094	0.284
30	0.094	0.345
50	0.100	0.383
100	0.096	0.434

Looking at the results above, we see that the Wilcoxon rank sum test has a noticeably smaller simulated type I error than the ordered comparisons test in all four cases and at all four sample sizes. The overall average increase in type I error for the ordered comparisons test was 0.259. Table 27 summarizes the average increase in simulated type I error for the ordered comparisons test across the various sample sizes analyzed.

Table 27. Average Simulated Type I Error Increase for the Ordered Comparisons Test over the Wilcoxon Rank Sum Test

Sample Size ($m = n = M$)	Average Increase in Type I Error
10	0.182
30	0.254
50	0.291
100	0.308

It is worth noting that both tests could be considered liberal in that they both had a higher type I error rate than the specified error rate, $\alpha \approx .05$.

With its increased power, the ordered comparisons test is more protective against type II errors (false negatives) and is more likely to pick up on differences between the two population distributions than the Wilcoxon rank sum test. However, this comes at the price of an increased type I error (false positives), and the ordered comparisons test is more likely to conclude that there is a difference between the two population distributions when in fact there may not be. Nevertheless, one could argue that this is a trade worth making considering that in many cases it is very unlikely for the two data samples to come from the exact same distribution and more likely that there exists at least a shift in population mean, small as it may be. In the case where we are interested in picking the best system, we typically assume that there is in fact a difference between the competing systems and we are not concerned with a type I error. In this case it would be more beneficial to use the ordered comparisons test with its superior ability at detecting differences in distributions, especially at small sample sizes and when the difference between competing systems is small.

3.6 Conclusions

We have proposed an effective nonparametric hypothesis test for determining if one distribution is stochastically larger (or smaller) than another. Our results show that the distribution-free ordered comparisons test is more powerful than the Wilcoxon rank sum test at detecting differences in distributions for all test cases analyzed. The improvement in performance is most noticeable at small sample sizes where nonparametric tests are most beneficial, as well as when there exist only small differences between the systems. The Wilcoxon rank sum test has smaller type I error on average. However, a sacrifice in type I error for the increase in power of the ordered comparisons test is of particular importance when testing whether one population is stochastically larger (or smaller) than another since it is rare for the sample data to come from the exact same population. By separately ranking the data from each sample and making ordered comparisons, our test makes full use of the information contained in the sample data.

IV. Conclusions

Throughout both papers, we have analyzed the critical failure data for the alternative preventative maintenance policies the Air Force is considering to increase the time between critical failures for an important Air Force system. We have shown various example calculations with both the newly proposed methods as well as several current analysis techniques. Table 28 below summarizes the results from each of the multinomial probability estimate methods discussed in Section II.

Table 28. Summary of Multinomial Probability Estimates for the Air Force's Preventative Maintenance Policy A ($\tilde{p}_1$) and Policy B ($\tilde{p}_2$)

Method	Multinomial Probability Estimate	
	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.3	0.7
AVC	0.33	0.67
Exact Integral	$1/3$ = 0.3333	$2/3$ = 0.6667
Kernel Density	0.3906	0.6094

In all techniques, the multinomial probability estimate for Policy B is roughly twice as large as the multinomial probability estimate for Policy A. This means that Policy B is twice as likely to result in a longer time between critical failures as Policy A. Therefore, we can recommend from this analysis that the Air Force should implement Policy B as a more effective preventative maintenance policy for increasing the time between critical failures.

In Section III, using the ordered comparisons hypothesis test, we concluded that we should reject H_0 (that there is no statistical difference between Policy A and Policy B). Therefore, we can say that there is enough statistical evidence at the $\alpha = .0547$ level of significance to conclude that $F_2(\cdot)$ is stochastically larger than $F_1(\cdot)$. That is, $\Pr\{\text{Policy B } \geq x\} \geq \Pr\{\text{Policy A } \geq x\}$, for all x, and for some x, $\Pr\{\text{Policy B } \geq x\} > \Pr\{\text{Policy A } \geq x\}$. Therefore, if current Air Force regulation states that systems should have a minimum time between critical failures of $x = 1.25$ months, we can say with 94.53% confidence that there is a greater probability that Policy B will fulfill this requirement than Policy A. In fact, we found an unbiased probability estimate, $\hat{\rho}$, that showed there is a 0.9 probability that Policy B is stochastically greater than Policy A.

It is worth noting that at the same level of significance, the Wilcoxon rank sum test fails to reject H_0 for the same set of failure time data. Since we know that the failure times for Policy B are from an exponential distribution with a mean time between critical failures of 2 months and the failure times for Policy A are from an exponential distribution with a mean time between critical failures of only 1 month, we know that there is in fact a difference between the two policies. Therefore, we can see that the ordered comparisons test is more powerful in picking up on smaller differences between systems than the Wilcoxon rank sum test.

Overall, we have presented two additional methods not currently in the literature for effectively comparing competing systems. The generalized integral method provides the ability to calculate the exact probability that a single system is best among competing systems in a single trial, when the analyst has knowledge of the distribution of the parameter of interest for each system. When this is the case, we can avoid calculating computationally intensive empirical probability estimates and save potentially large amounts of time and money attempting to quantify differences between systems. However, if distributional knowledge of the systems is unknown, we also showed that there are still several techniques which can be used to calculate accurate empirical probability estimates using the generalized integral method.

While these empirical probability estimation techniques provide sufficient results, they are more computationally intensive than current empirical techniques. Further analysis ought to be conducted in order uncover a more efficient and accurate procedure for estimating the pdf and cdf of the parameter of interest for each system. We also briefly explored a bootstrapping density estimation technique that was easy to employ, however, accurate multinomial probability estimates were only achieved when using a large number of bootstrap samples. Results from that analysis exhibited probability estimate accuracies that were comparable with that of Procedure AVC, but again, a more thorough analysis should be conducted in order to find an improved empirical density estimation procedure.

In theory, the kernel density estimation technique is appealing, however, in practice the results left something to be desired. A further analysis should be conducted in order to determine optimal settings for the bandwidth parameter, h, as well as the appropriate kernel to use for specific situations. A more in depth search of density estimation techniques may also uncover a more effective method for modeling density functions.

The nonparametric ordered comparisons test proved very effective in determining if one system is stochastically larger than another. Although it makes a slight sacrifice in type I error, the ordered comparisons test is far more effective than the Wilcoxon rank sum test at detecting differences between systems. The improvement in performance was most noticeable at small sample sizes and when the true difference between the two systems was small. At a sample size of $M = 10$, the ordered comparison test was 81.68% more powerful than the Wilcoxon rank sum test. Likewise, for an indifference parameter value of $\theta = 1.2$, the ordered comparisons test was 204.3% more powerful than the Wilcoxon rank sum test. By separately ranking the data from each sample and making ordered comparisons, our test makes full use of the information contained in the sample data.

When trying to determine the best system among competing systems, it is typical to consider a system as best if it has the largest mean or best long run performance. However, we have shown that there are cases when this determination should be based on one-time performance. There are also instances when long run average performance data is unavailable or we are dealing with only a small dataset but would still like to determine which system is best. The generalized integral method and ordered comparisons hypothesis test have proven to be effective alternative approaches for achieving this goal.

Appendix A: Empirical Probability Estimate and Error Summary Tables

Table 29. Empirical Probability Estimates for Test Case 1: Beta(α = 1,β = 3) vs. Beta(α = 1,β = 2.5), θ = 1.2

	Sample Size (v)						Average Probabilities	
Method	10		30		100			
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.4589	0.5411	0.4567	0.5433	0.4530	0.5470	0.4562	0.5438
AVC	0.4586	0.5414	0.4556	0.5444	0.4534	0.5466	0.4559	0.5441
Kernel Density	0.4612	0.5381	0.4589	0.5407	0.4555	0.5440	0.4585	0.5410

Table 30. Empirical Probability Estimate Errors for Test Case 1: Beta(α = 1,β = 3) vs. Beta(α = 1,β = 2.5), θ = 1.2

	Sample Size (v)						Overall Average Error	
Method	10		30		100			
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0043	0.0043	0.0022	0.0022	0.0016	0.0016	0.0027	0.0027
AVC	0.0041	0.0041	0.0011	0.0011	0.0011	0.0011	0.0021	0.0021
Kernel Density	0.0066	0.0073	0.0043	0.0048	0.0009	0.0014	0.0040	0.0045

Table 31. Empirical Probability Estimates for Test Case 2: Beta(α = 2,β = 5) vs. Beta(α = 2,β = 2), θ = 3.4

	Sample Size (v)						Average Probabilities	
Method	10		30		100			
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.2224	0.7776	0.2232	0.7768	0.2263	0.7737	0.2240	0.7760
AVC	0.2210	0.7790	0.2234	0.7766	0.2260	0.7740	0.2235	0.7765
Kernel Density	0.2435	0.7563	0.2405	0.7591	0.2374	0.7626	0.2405	0.7593

Table 32. Empirical Probability Estimate Errors for Test Case 2: Beta(α = 2,β = 5) vs. Beta(α = 2,β = 2), θ = 3.4

	Sample Size (v)						Overall Average Error	
Method	10		30		100			
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0038	0.0038	0.0030	0.0030	0.0001	0.0001	0.0023	0.0023
AVC	0.0052	0.0052	0.0028	0.0028	0.0002	0.0002	0.0027	0.0027
Kernel Density	0.0173	0.0175	0.0143	0.0147	0.0112	0.0112	0.0143	0.0145

Table 33. Empirical Probability Estimates for Test Case 3: Exponential(μ = 1) vs. Exponential(μ = 1.2), $\boldsymbol{\theta}$ = 1.2

Method	Sample Size (v) 10		30		100		Average Probabilities	
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.4564	0.5436	0.4500	0.5500	0.4539	0.5461	0.4535	0.5465
AVC	0.4573	0.5427	0.4508	0.5492	0.4535	0.5465	0.4539	0.5461
Kernel Density	0.4616	0.5384	0.4542	0.5457	0.4560	0.5439	0.4573	0.5427

Table 34. Empirical Probability Estimate Errors for Test Case 3: Exponential(μ = 1) vs. Exponential(μ = 1.2), $\boldsymbol{\theta}$ = 1.2

Method	Sample Size (v) 10		30		100		Overall Average Error	
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0019	0.0019	0.0045	0.0045	0.0006	0.0006	0.0024	0.0024
AVC	0.0028	0.0028	0.0038	0.0038	0.0010	0.0010	0.0025	0.0025
Kernel Density	0.0071	0.0071	0.0003	0.0003	0.0015	0.0015	0.0030	0.0030

Table 35. Empirical Probability Estimates for Test Case 4: Exponential(μ = 1) vs. Exponential(μ = 2.2), $\boldsymbol{\theta}$ = 2.2

Method	Sample Size (v) 10		30		100		Average Probabilities	
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.3076	0.6924	0.3147	0.6853	0.3113	0.6887	0.3112	0.6888
AVC	0.3122	0.6878	0.3125	0.6875	0.3121	0.6879	0.3123	0.6877
Kernel Density	0.3315	0.6685	0.3299	0.6701	0.3250	0.6750	0.3288	0.6712

Table 36. Empirical Probability Estimate Errors for Test Case 4: Exponential(μ = 1) vs. Exponential(μ = 2.2), $\boldsymbol{\theta}$ = 2.2

Method	Sample Size (v) 10		30		100		Overall Average Error	
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0049	0.0049	0.0022	0.0022	0.0012	0.0012	0.0028	0.0028
AVC	0.0003	0.0003	0.0000	0.0000	0.0004	0.0004	0.0002	0.0002
Kernel Density	0.0190	0.0190	0.0174	0.0174	0.0125	0.0125	0.0163	0.0163

Table 37. Empirical Probability Estimates for Test Case 5: Exponential(μ = 1) vs. Exponential(μ = 3.4), θ = 3.4

	Sample Size (v)						Average Probabilities	
Method	10		30		100			
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.2257	0.7743	0.2269	0.7731	0.2259	0.7741	0.2262	0.7738
AVC	0.2250	0.7750	0.2267	0.7733	0.2257	0.7743	0.2258	0.7742
Kernel Density	0.2594	0.7406	0.2570	0.7430	0.2485	0.7514	0.2550	0.7450

Table 38. Empirical Probability Estimate Errors for Test Case 5: Exponential(μ = 1) vs. Exponential(μ = 3.4), θ = 3.4

	Sample Size (v)						Overall Average Error	
Method	10		30		100			
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0016	0.0016	0.0004	0.0004	0.0014	0.0014	0.0011	0.0011
AVC	0.0022	0.0022	0.0006	0.0006	0.0016	0.0016	0.0015	0.0015
Kernel Density	0.0322	0.0322	0.0297	0.0297	0.0212	0.0214	0.0277	0.0277

Table 39. Empirical Probability Estimates for Test Case 6: Gamma(α = 2,β = 2) vs. Gamma(α = 3,β = 2), θ = 2.2

	Sample Size (v)						Average Probabilities	
Method	10		30		100			
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.3063	0.6937	0.3150	0.6850	0.3133	0.6867	0.3115	0.6885
AVC	0.3092	0.6908	0.3165	0.6835	0.3149	0.6851	0.3135	0.6865
Kernel Density	0.3316	0.6684	0.3347	0.6653	0.3281	0.6719	0.3314	0.6686

Table 40. Empirical Probability Estimate Errors for Test Case 6: Gamma(α = 2,β = 2) vs. Gamma(α = 3,β = 2), θ = 2.2

	Sample Size (v)						Overall Average Error	
Method	10		30		100			
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0062	0.0062	0.0025	0.0025	0.0008	0.0008	0.0031	0.0031
AVC	0.0033	0.0033	0.0040	0.0040	0.0024	0.0024	0.0032	0.0032
Kernel Density	0.0191	0.0191	0.0222	0.0222	0.0156	0.0156	0.0189	0.0189

Table 41. Empirical Probability Estimates for
Test Case 7: Normal($\mu = 0, \sigma = 1$) vs. Normal($\mu = 0.5, \sigma = 1$), θ = 1.8

Method	Sample Size (v)						Average Probabilities	
	10		30		100			
	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_1$	$\tilde{p}_2$
BEM	0.3680	0.6320	0.3615	0.6385	0.3606	0.6394	0.3634	0.6366
AVC	0.3626	0.6374	0.3617	0.6383	0.3620	0.6380	0.3621	0.6379
Kernel Density	0.3754	0.6246	0.3720	0.6280	0.3692	0.6308	0.3722	0.6278

Table 42. Empirical Probability Estimate Errors for
Test Case 7: Normal(μ = 0,σ = 1) vs. Normal(μ = 0.5,σ = 1), θ = 1.8

Method	Sample Size (v)						Overall Average Error	
	10		30		100			
	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\varepsilon_{\tilde{p}_1}$	$\varepsilon_{\tilde{p}_2}$	$\bar{\varepsilon}_{\tilde{p}_1}$	$\bar{\varepsilon}_{\tilde{p}_2}$
BEM	0.0061	0.0061	0.0003	0.0003	0.0012	0.0012	0.0025	0.0025
AVC	0.0008	0.0008	0.0001	0.0001	0.0001	0.0001	0.0004	0.0004
Kernel Density	0.0135	0.0135	0.0101	0.0101	0.0074	0.0074	0.0103	0.0103

Appendix B: Ordered Comparisons Hypothesis Test Probability Tables

Table 43. Upper-Tail Probabilities for the Null Distribution of the Ordered Comparisons Test

For a given M, the table entry for the point ℓ is $P_0\{L \geq x\}$. Under these conditions, if x is such that $P_0\{L \geq x\} = \alpha$, then $\ell_\alpha = x$.

					M					
x	2	3	4	5	6	7	8	9	10	11
1	.7500									
2	.2500	.5000	.6875							
3		.1250	.3125	.5000	.6563					
4			.0625	.1875	.3438	.5000	.6367			
5				.0313	.1094	.2266	.3633	.5000	.6230	
6					.0156	.0625	.1445	.2539	.3770	.5000
7						.0078	.0352	.0898	.1719	.2744
8							.0039	.0195	.0547	.1133
9								.0020	.0107	.0327
10									.0010	.0059
11										.0005

					M					
x	12	13	14	15	16	17	18	19	20	21
6	.6128									
7	.3872	.5000	.6047							
8	.1938	.2905	.3953	.5000	.5982					
9	.0730	.1334	.2120	.3036	.4018	.5000	.5927			
10	.0193	.0461	.0898	.1509	.2272	.3145	.4073	.5000	.5881	
11	.0032	.0112	.0287	.0592	.1051	.1662	.2403	.3238	.4119	.5000
12	.0002	.0017	.0065	.0176	.0384	.0717	.1189	.1796	.2517	.3318
13		.0001	.0009	.0037	.0106	.0245	.0481	.0835	.1316	.1917
14			.0001	.0005	.0021	.0064	.0154	.0318	.0577	.0946
15				.0000	.0003	.0012	.0038	.0096	.0207	.0392
16					.0000	.0001	.0007	.0022	.0059	.0133
17						.0000	.0001	.0004	.0013	.0036
18							.0000	.0000	.0002	.0007
19								.0000	.0000	.0001
20									.0000	.0000
21										.0000

					M					
x	22	23	24	25	26	27	28	29	30	31
11	.5841									
12	.4159	.5000	.5806							
13	.2617	.3388	.4194	.5000	.5775					
14	.1431	.2024	.2706	.3450	.4225	.5000	.5747			
15	.0669	.1050	.1537	.2122	.2786	.3506	.4253	.5000	.5722	
16	.0262	.0466	.0758	.1148	.1635	.2210	.2858	.3555	.4278	.5000
17	.0085	.0173	.0320	.0539	.0843	.1239	.1725	.2291	.2923	.3601
18	.0022	.0053	.0113	.0216	.0378	.0610	.0925	.1325	.1808	.2366
19	.0004	.0013	.0033	.0073	.0145	.0261	.0436	.0680	.1002	.1405
20	.0001	.0002	.0008	.0020	.0047	.0096	.0178	.0307	.0494	.0748
21	.0000	.0000	.0001	.0005	.0012	.0030	.0063	.0121	.0214	.0354
22	.0000	.0000	.0000	.0001	.0003	.0008	.0019	.0041	.0081	.0147
23		.0000	.0000	.0000	.0000	.0002	.0005	.0012	.0026	.0053
24			.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0017
25				.0000	.0000	.0000	.0000	.0001	.0002	.0004
26					.0000	.0000	.0000	.0000	.0000	.0001
27						.0000	.0000	.0000	.0000	.0000
28							.0000	.0000	.0000	.0000
29								.0000	.0000	.0000
30									.0000	.0000
31										.0000

					M					
x	32	33	34	35	36	37	38	39	40	41
16	.5700									
17	.4300	.5000	.5679							
18	.2983	.3642	.4321	.5000	.5660					
19	.1885	.2434	.3038	.3679	.4340	.5000	.5643			
20	.1077	.1481	.1958	.2498	.3089	.3714	.4357	.5000	.5627	
21	.0551	.0814	.1147	.1553	.2025	.2557	.3136	.3746	.4373	.5000
22	.0251	.0401	.0607	.0877	.1215	.1620	.2088	.2612	.3179	.3776
23	.0100	.0175	.0288	.0448	.0662	.0939	.1279	.1684	.2148	.2664
24	.0035	.0068	.0122	.0205	.0326	.0494	.0717	.0998	.1341	.1744
25	.0011	.0023	.0045	.0083	.0144	.0235	.0365	.0541	.0769	.1055
26	.0003	.0007	.0015	.0030	.0057	.0100	.0168	.0266	.0403	.0586
27	.0001	.0002	.0004	.0009	.0020	.0038	.0069	.0119	.0192	.0298
28	.0000	.0000	.0001	.0003	.0006	.0013	.0025	.0047	.0083	.0138
29	.0000	.0000	.0000	.0001	.0002	.0004	.0008	.0017	.0032	.0058
30	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0011	.0022
31	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0007
32	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
33		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
34			.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
35				.0000	.0000	.0000	.0000	.0000	.0000	.0000
36					.0000	.0000	.0000	.0000	.0000	.0000
37						.0000	.0000	.0000	.0000	.0000
38							.0000	.0000	.0000	.0000
39								.0000	.0000	.0000
40									.0000	.0000
41										.0000

	M									
x	42	43	44	45	46	47	48	49	50	51
21	.5612									
22	.4388	.5000	.5598							
23	.3220	.3804	.4402	.5000	.5585					
24	.2204	.2712	.3258	.3830	.4415	.5000	.5573			
25	.1400	.1802	.2257	.2757	.3294	.3854	.4427	.5000	.5561	
26	.0821	.1110	.1456	.1856	.2307	.2800	.3327	.3877	.4439	.5000
27	.0442	.0631	.0871	.1163	.1510	.1908	.2354	.2841	.3359	.3899
28	.0218	.0330	.0481	.0676	.0920	.1215	.1562	.1958	.2399	.2879
29	.0098	.0158	.0244	.0362	.0519	.0719	.0967	.1264	.1611	.2005
30	.0040	.0069	.0113	.0178	.0270	.0395	.0557	.0762	.1013	.1312
31	.0014	.0027	.0048	.0080	.0129	.0200	.0297	.0427	.0595	.0804
32	.0005	.0010	.0018	.0033	.0057	.0093	.0147	.0222	.0325	.0460
33	.0001	.0003	.0006	.0012	.0023	.0040	.0066	.0106	.0164	.0244
34	.0000	.0001	.0002	.0004	.0008	.0015	.0028	.0047	.0077	.0120
35	.0000	.0000	.0001	.0001	.0003	.0005	.0010	.0019	.0033	.0055
36	.0000	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0013	.0023
37	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0009
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
40	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
41	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
42	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
43		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
44			.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
45				.0000	.0000	.0000	.0000	.0000	.0000	.0000
46					.0000	.0000	.0000	.0000	.0000	.0000
47						.0000	.0000	.0000	.0000	.0000
48							.0000	.0000	.0000	.0000
49								.0000	.0000	.0000
50									.0000	.0000
51										.0000

					M					
x	52	53	54	55	56	57	58	59	60	61
26	0.5551									
27	0.4449	0.5000	0.5540							
28	0.3389	0.3919	0.4460	0.5000	0.5531					
29	0.2442	0.2916	0.3417	0.3939	0.4469	0.5000	0.5522			
30	0.1659	0.2051	0.2483	0.2950	0.3444	0.3957	0.4478	0.5000	0.5513	
31	0.1058	0.1358	0.1704	0.2094	0.2522	0.2983	0.3470	0.3974	0.4487	0.5000
32	0.0632	0.0845	0.1102	0.1403	0.1748	0.2135	0.2559	0.3015	0.3494	0.3991
33	0.0352	0.0492	0.0668	0.0885	0.1144	0.1446	0.1791	0.2175	0.2595	0.3045
34	0.0182	0.0267	0.0380	0.0524	0.0704	0.0924	0.1185	0.1488	0.1831	0.2213
35	0.0088	0.0135	0.0201	0.0290	0.0407	0.0556	0.0740	0.0963	0.1225	0.1528
36	0.0039	0.0063	0.0099	0.0150	0.0220	0.0314	0.0435	0.0587	0.0775	0.1000
37	0.0016	0.0027	0.0045	0.0072	0.0111	0.0166	0.0240	0.0337	0.0462	0.0619
38	0.0006	0.0011	0.0019	0.0032	0.0052	0.0082	0.0124	0.0182	0.0259	0.0361
39	0.0002	0.0004	0.0007	0.0013	0.0023	0.0038	0.0060	0.0092	0.0137	0.0198
40	0.0001	0.0001	0.0003	0.0005	0.0009	0.0016	0.0027	0.0043	0.0067	0.0102
41	0.0000	0.0000	0.0001	0.0002	0.0003	0.0006	0.0011	0.0019	0.0031	0.0049
42	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004	0.0008	0.0013	0.0022
43	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0009
44	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004
45	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
46	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
47	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
48	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
49	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
51	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
52	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
53		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
54			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
55				0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
56					0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
57						0.0000	0.0000	0.0000	0.0000	0.0000
58							0.0000	0.0000	0.0000	0.0000
59								0.0000	0.0000	0.0000
60									0.0000	0.0000
61										0.0000

	M									
x	62	63	64	65	66	67	68	69	70	71
31	0.5505									
32	0.4495	0.5000	0.5497							
33	0.3518	0.4007	0.4503	0.5000	0.5489					
34	0.2629	0.3073	0.3540	0.4022	0.4511	0.5000	0.5482			
35	0.1871	0.2250	0.2662	0.3101	0.3561	0.4036	0.4518	0.5000	0.5475	
36	0.1264	0.1568	0.1909	0.2285	0.2693	0.3127	0.3582	0.4050	0.4525	0.5000
37	0.0809	0.1037	0.1302	0.1605	0.1945	0.2319	0.2723	0.3152	0.3601	0.4063
38	0.0490	0.0650	0.0843	0.1073	0.1339	0.1642	0.1981	0.2352	0.2752	0.3177
39	0.0279	0.0385	0.0517	0.0680	0.0876	0.1108	0.1375	0.1678	0.2015	0.2383
40	0.0150	0.0215	0.0300	0.0408	0.0544	0.0710	0.0909	0.1142	0.1410	0.1712
41	0.0076	0.0113	0.0164	0.0232	0.0320	0.0432	0.0571	0.0740	0.0941	0.1175
42	0.0036	0.0056	0.0084	0.0124	0.0178	0.0249	0.0341	0.0456	0.0598	0.0770
43	0.0016	0.0026	0.0041	0.0063	0.0093	0.0136	0.0192	0.0266	0.0361	0.0480
44	0.0006	0.0011	0.0018	0.0030	0.0046	0.0070	0.0103	0.0147	0.0207	0.0284
45	0.0002	0.0004	0.0008	0.0013	0.0021	0.0034	0.0052	0.0077	0.0112	0.0160
46	0.0001	0.0002	0.0003	0.0005	0.0009	0.0015	0.0025	0.0038	0.0058	0.0085
47	0.0000	0.0001	0.0001	0.0002	0.0004	0.0007	0.0011	0.0018	0.0028	0.0043
48	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0005	0.0008	0.0013	0.0020
49	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0009
50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004
51	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
52	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
53	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
54	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
55	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
56	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
57	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
58	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
61	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
62	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
63		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
65				0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
66					0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
67						0.0000	0.0000	0.0000	0.0000	0.0000
68							0.0000	0.0000	0.0000	0.0000
69								0.0000	0.0000	0.0000
70									0.0000	0.0000
71										0.0000

	M									
x	72	73	74	75	76	77	78	79	80	81
36	0.5469									
37	0.4531	0.5000	0.5462							
38	0.3620	0.4076	0.4538	0.5000	0.5456					
39	0.2780	0.3200	0.3638	0.4088	0.4544	0.5000	0.5450			
40	0.2048	0.2414	0.2807	0.3222	0.3655	0.4099	0.4550	0.5000	0.5445	
41	0.1444	0.1746	0.2080	0.2443	0.2833	0.3244	0.3672	0.4111	0.4555	0.5000
42	0.0973	0.1208	0.1477	0.1778	0.2111	0.2472	0.2858	0.3265	0.3688	0.4122
43	0.0625	0.0799	0.1003	0.1240	0.1509	0.1810	0.2141	0.2499	0.2882	0.3285
44	0.0382	0.0503	0.0651	0.0827	0.1034	0.1272	0.1541	0.1841	0.2170	0.2526
45	0.0222	0.0302	0.0403	0.0527	0.0677	0.0855	0.1063	0.1302	0.1572	0.1871
46	0.0122	0.0172	0.0237	0.0320	0.0423	0.0550	0.0703	0.0883	0.1093	0.1332
47	0.0064	0.0093	0.0133	0.0185	0.0252	0.0338	0.0444	0.0573	0.0728	0.0910
48	0.0032	0.0048	0.0070	0.0101	0.0143	0.0198	0.0268	0.0356	0.0465	0.0596
49	0.0015	0.0023	0.0035	0.0053	0.0077	0.0110	0.0154	0.0211	0.0283	0.0374
50	0.0006	0.0011	0.0017	0.0026	0.0040	0.0058	0.0084	0.0119	0.0165	0.0224
51	0.0003	0.0005	0.0008	0.0012	0.0019	0.0029	0.0044	0.0064	0.0092	0.0128
52	0.0001	0.0002	0.0003	0.0005	0.0009	0.0014	0.0022	0.0033	0.0048	0.0070
53	0.0000	0.0001	0.0001	0.0002	0.0004	0.0006	0.0010	0.0016	0.0024	0.0036
54	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0007	0.0012	0.0018
55	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0008
56	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004
57	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
58	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
61	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
62	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
65	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
66	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
67	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
68	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
69	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
70	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
73		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
74			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
75				0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
76					0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
77						0.0000	0.0000	0.0000	0.0000	0.0000
78							0.0000	0.0000	0.0000	0.0000
79								0.0000	0.0000	0.0000
80									0.0000	0.0000
81										0.0000

					M					
x	82	83	84	85	86	87	88	89	90	91
41	0.5439									
42	0.4561	0.5000	0.5434							
43	0.3703	0.4132	0.4566	0.5000	0.5429					
44	0.2906	0.3304	0.3718	0.4142	0.4571	0.5000	0.5424			
45	0.2199	0.2552	0.2928	0.3323	0.3733	0.4152	0.4576	0.5000	0.5419	
46	0.1601	0.1900	0.2226	0.2577	0.2950	0.3341	0.3747	0.4161	0.4581	0.5000
47	0.1121	0.1361	0.1631	0.1928	0.2253	0.2601	0.2971	0.3359	0.3760	0.4170
48	0.0753	0.0937	0.1149	0.1390	0.1659	0.1956	0.2279	0.2625	0.2992	0.3376
49	0.0485	0.0619	0.0778	0.0964	0.1177	0.1418	0.1687	0.1983	0.2304	0.2648
50	0.0299	0.0392	0.0506	0.0642	0.0803	0.0990	0.1204	0.1445	0.1714	0.2009
51	0.0176	0.0238	0.0315	0.0410	0.0526	0.0665	0.0827	0.1016	0.1231	0.1472
52	0.0099	0.0138	0.0188	0.0251	0.0331	0.0428	0.0546	0.0687	0.0851	0.1041
53	0.0053	0.0076	0.0107	0.0147	0.0199	0.0265	0.0347	0.0447	0.0567	0.0709
54	0.0027	0.0040	0.0058	0.0083	0.0115	0.0157	0.0211	0.0279	0.0363	0.0465
55	0.0013	0.0020	0.0030	0.0044	0.0063	0.0089	0.0123	0.0167	0.0223	0.0293
56	0.0006	0.0010	0.0015	0.0023	0.0033	0.0048	0.0069	0.0096	0.0132	0.0177
57	0.0003	0.0004	0.0007	0.0011	0.0017	0.0025	0.0037	0.0053	0.0074	0.0103
58	0.0001	0.0002	0.0003	0.0005	0.0008	0.0012	0.0019	0.0028	0.0040	0.0057
59	0.0000	0.0001	0.0001	0.0002	0.0004	0.0006	0.0009	0.0014	0.0021	0.0031
60	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0007	0.0010	0.0016
61	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0008
62	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
64	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
65	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
66	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
67	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
68	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
69	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
70	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
73	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
74	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
76	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
77	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
78	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
79	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
80	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
81	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
82	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
83		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
85				0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
86					0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
87						0.0000	0.0000	0.0000	0.0000	0.0000
88							0.0000	0.0000	0.0000	0.0000
89								0.0000	0.0000	0.0000
90									0.0000	0.0000
91										0.0000

	M									
x	92	93	94	95	96	97	98	99	100	101
46	0.5415	0.5821	0.6214	0.6591	0.6950	0.7287	0.7602	0.7892	0.8159	0.8401
47	0.4585	0.5000	0.5410	0.5812	0.6202	0.6576	0.6931	0.7266	0.7579	0.7869
48	0.3773	0.4179	0.4590	0.5000	0.5406	0.5804	0.6190	0.6561	0.6914	0.7246
49	0.3012	0.3393	0.3786	0.4188	0.4594	0.5000	0.5402	0.5796	0.6178	0.6546
50	0.2329	0.2670	0.3031	0.3409	0.3798	0.4196	0.4598	0.5000	0.5398	0.5788
51	0.1741	0.2035	0.2352	0.2692	0.3050	0.3424	0.3810	0.4204	0.4602	0.5000
52	0.1257	0.1499	0.1767	0.2060	0.2376	0.2713	0.3069	0.3439	0.3822	0.4212
53	0.0875	0.1066	0.1282	0.1524	0.1792	0.2084	0.2398	0.2734	0.3086	0.3454
54	0.0587	0.0731	0.0898	0.1090	0.1307	0.1550	0.1817	0.2108	0.2421	0.2754
55	0.0379	0.0483	0.0607	0.0753	0.0921	0.1114	0.1332	0.1574	0.1841	0.2131
56	0.0235	0.0307	0.0395	0.0501	0.0627	0.0774	0.0944	0.1138	0.1356	0.1599
57	0.0140	0.0188	0.0247	0.0321	0.0411	0.0519	0.0646	0.0795	0.0967	0.1162
58	0.0080	0.0110	0.0149	0.0198	0.0260	0.0335	0.0427	0.0537	0.0666	0.0816
59	0.0044	0.0062	0.0086	0.0117	0.0158	0.0209	0.0272	0.0350	0.0443	0.0555
60	0.0023	0.0033	0.0048	0.0067	0.0092	0.0125	0.0167	0.0219	0.0284	0.0364
61	0.0012	0.0017	0.0025	0.0037	0.0052	0.0072	0.0098	0.0133	0.0176	0.0230
62	0.0006	0.0009	0.0013	0.0019	0.0028	0.0040	0.0056	0.0077	0.0105	0.0140
63	0.0003	0.0004	0.0006	0.0010	0.0014	0.0021	0.0030	0.0043	0.0060	0.0083
64	0.0001	0.0002	0.0003	0.0005	0.0007	0.0011	0.0016	0.0023	0.0033	0.0047
65	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0008	0.0012	0.0018	0.0025
66	0.0000	0.0000	0.0001	0.0001	0.0002	0.0002	0.0004	0.0006	0.0009	0.0013
67	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0007
68	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003
69	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001
70	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
73	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
74	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
76	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
77	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
78	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
79	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
80	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
81	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
82	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
85	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
86	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
82	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
85	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
86	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
87	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
88		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
89			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90				0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
⋮										⋮
101										0.0000

Appendix C: Simulated Power and Type I Error Plots

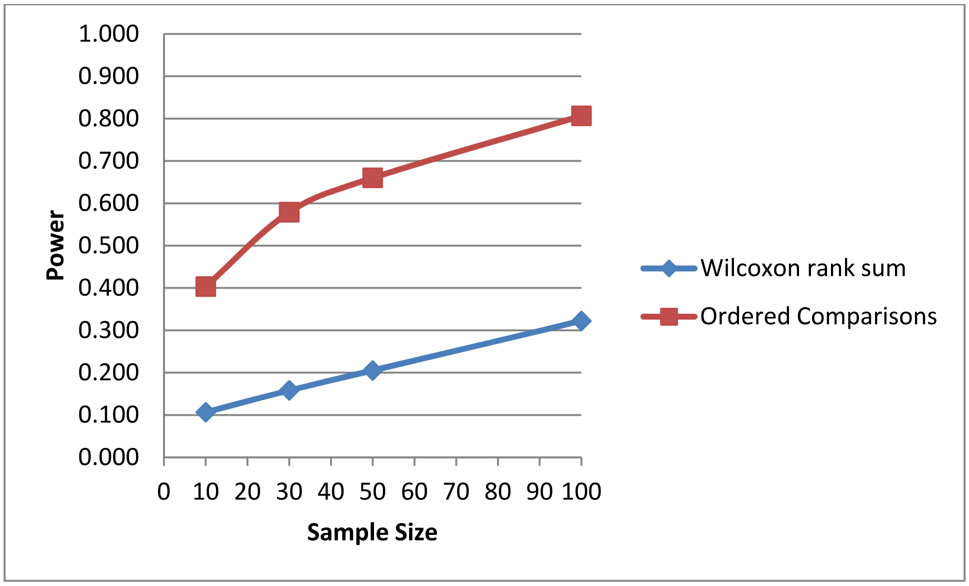

Figure 11. Simulated Power ($\alpha \approx .05$) for Test Case 1: Beta($\alpha = 1, \beta = 3$) vs. Beta($\alpha = 1, \beta = 2.5$), $\theta = 1.2$

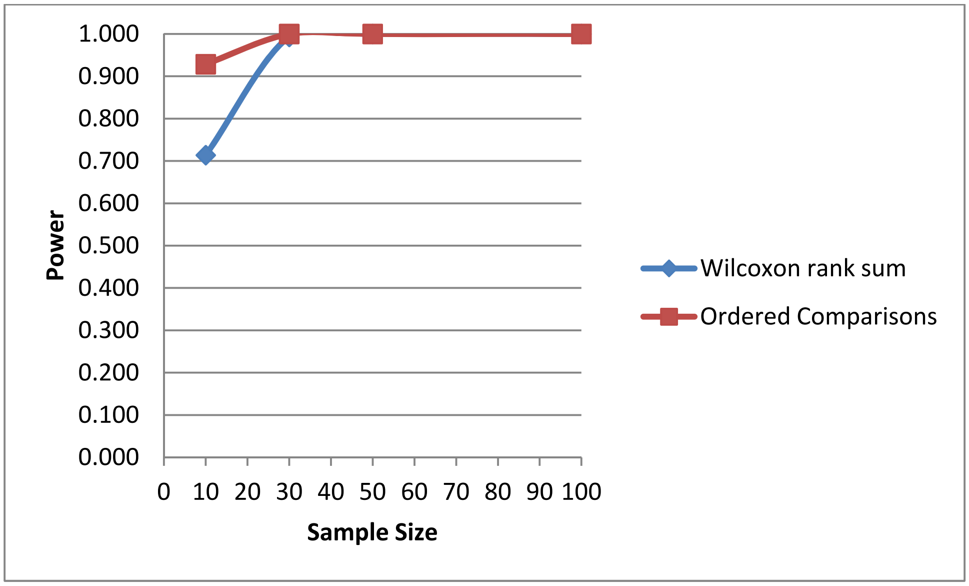

Figure 12. Simulated Power ($\alpha \approx .05$) for Test Case 2: Beta($\alpha = 2, \beta = 5$) vs. Beta($\alpha = 2, \beta = 2$), $\theta = 3.4$

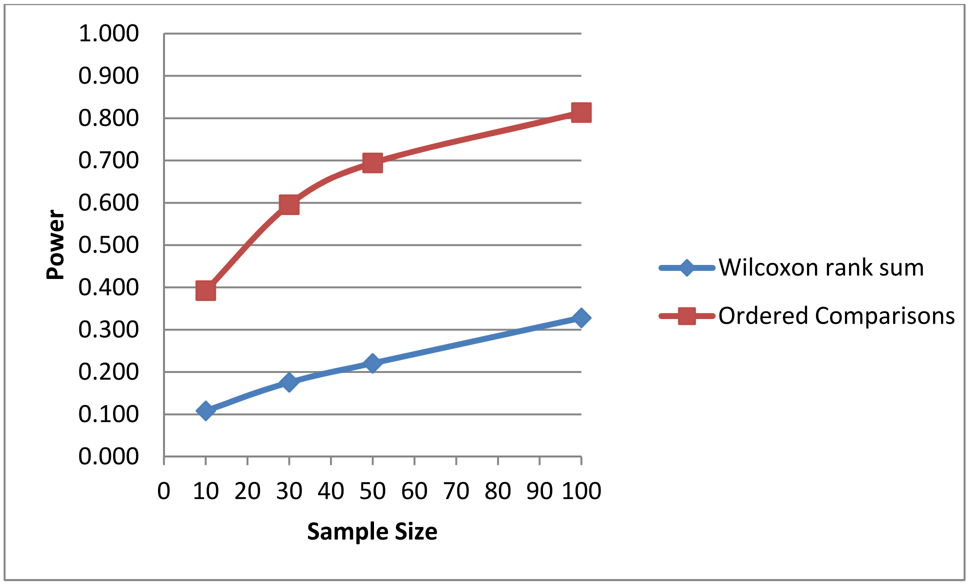

Figure 13. Simulated Power ($\alpha \approx .05$) for
Test Case 3: Exponential ($\mu = 1$) vs. Exponential ($\mu = 1.2$), θ = 1.2

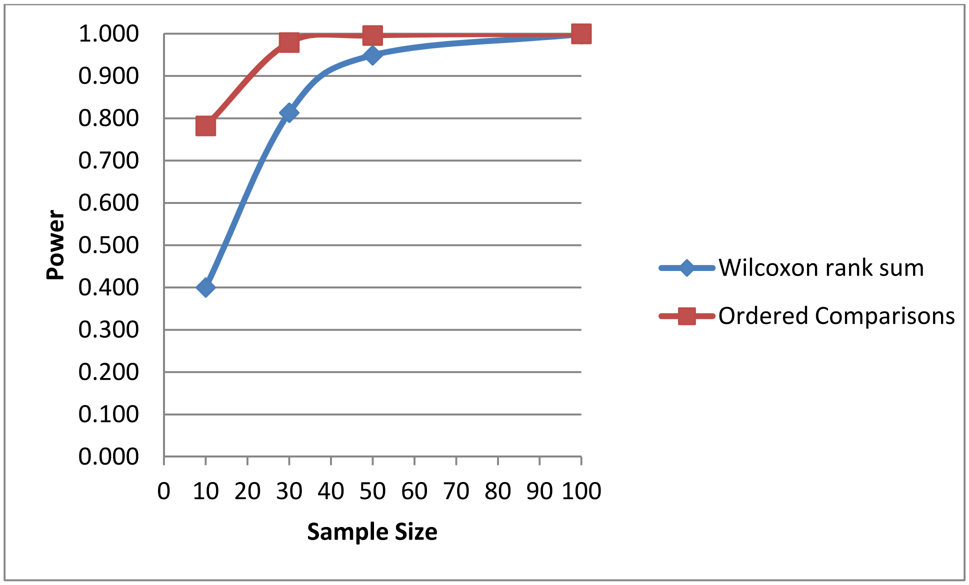

Figure 14. Simulated Power ($\alpha \approx .05$) for
Test Case 4: Exponential (μ = 1) vs. Exponential (μ = 2.2), θ = 2.2

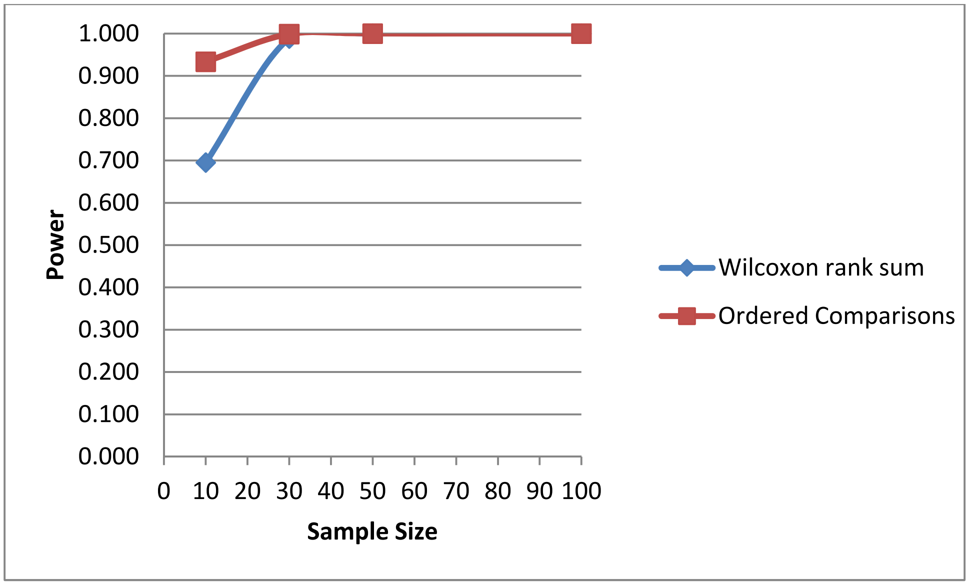

Figure 15. Simulated Power ($\alpha \approx .05$) for Test Case 5: Exponential ($\mu = 1$) vs. Exponential ($\mu = 3.4$), θ = 3.4

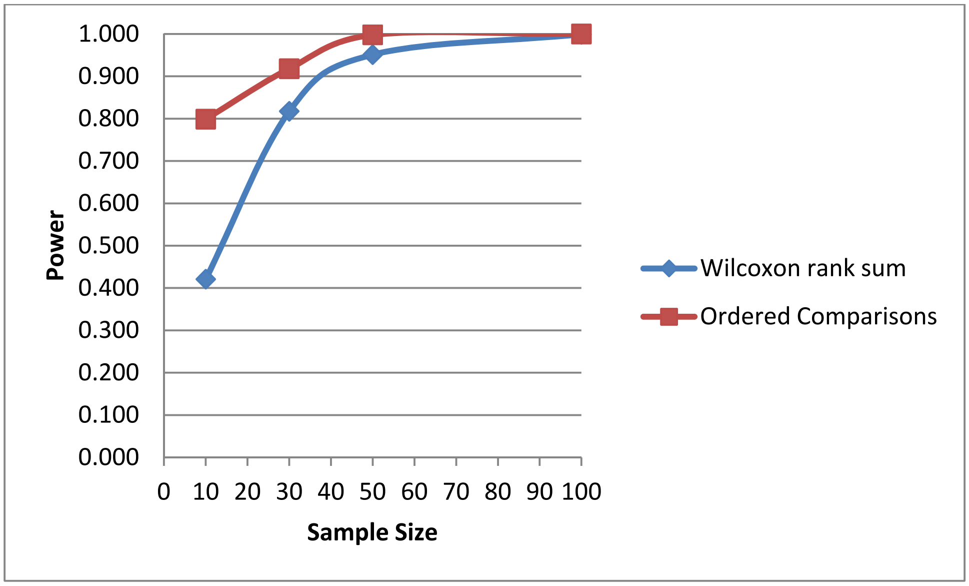

Figure 16. Simulated Power ($\alpha \approx .05$) for Test Case 6: Gamma ($\alpha = 2, \beta = 2$) vs. Gamma ($\alpha = 3, \beta = 2$), θ = 2.2

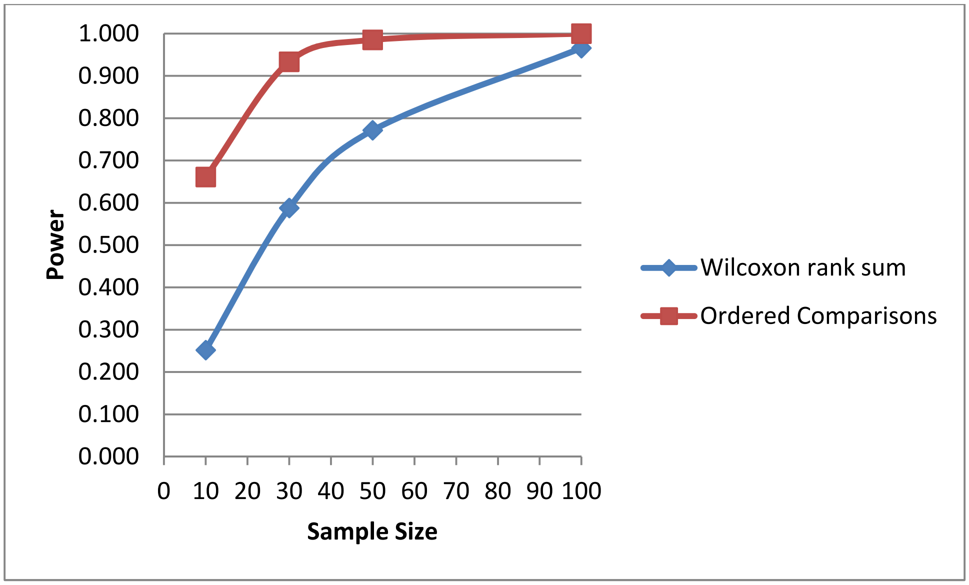

Figure 17. Simulated Power ($\alpha \approx .05$) for Test Case 7: Normal ($\mu = 0, \sigma = 1$) vs. Normal ($\mu = 0.5, \sigma = 1$), $\theta = 1.8$

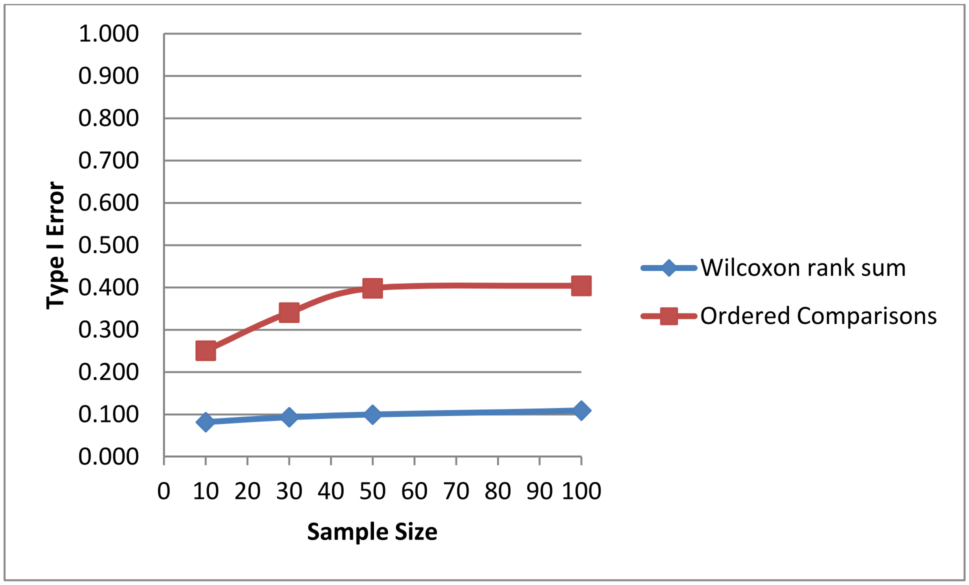

Figure 18. Simulated Type I Error for ($\alpha \approx .05$) for Beta ($\alpha = 1, \beta = 3$) vs. Beta ($\alpha = 1, \beta = 3$)

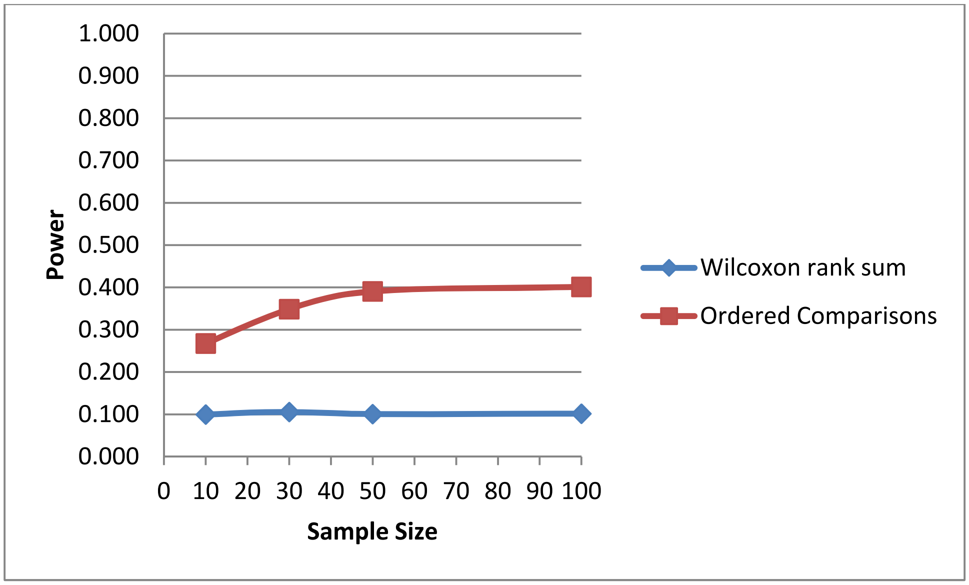

Figure 19. Simulated Type I Error ($\alpha \approx .05$) for Exponential ($\mu = 2$) vs. Exponential ($\mu = 2$)

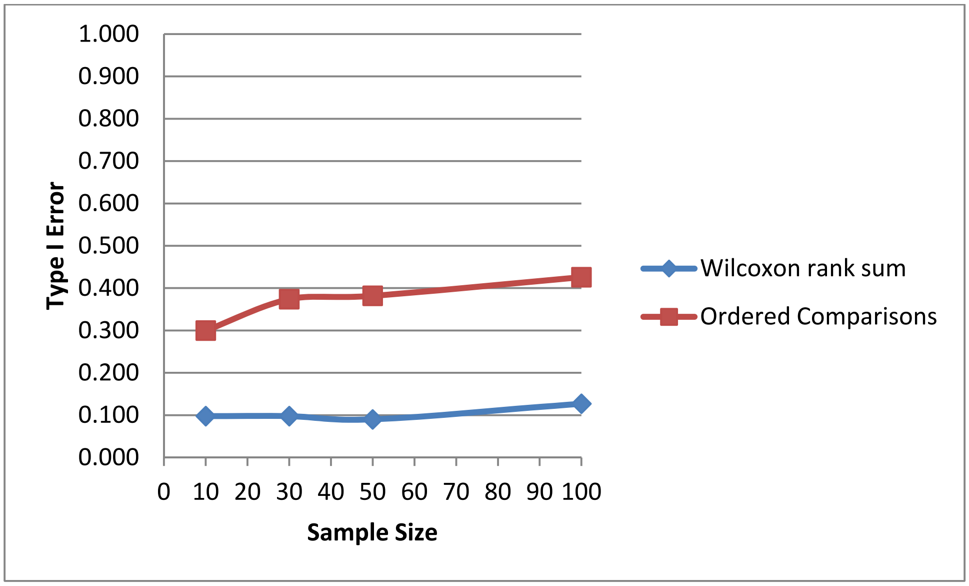

Figure 20. Simulated Type I Error ($\alpha \approx .05$) for Gamma ($\alpha = 2, \beta = 2$) vs. Gamma ($\alpha = 2, \beta = 2$)

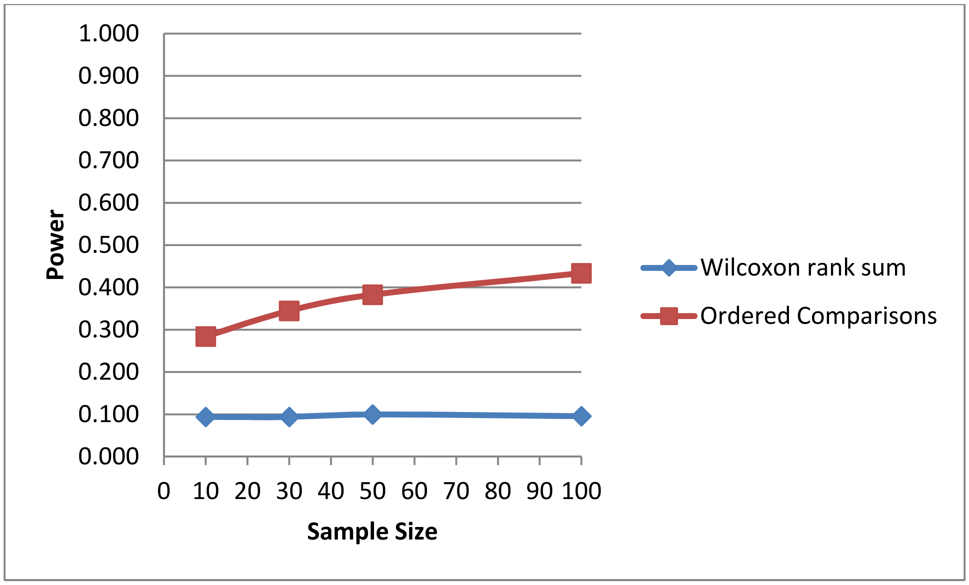

Figure 21. Simulated Type I Error ($\alpha \approx .05$) for Normal ($\mu = 0, \sigma = 1$) vs. Normal ($\mu = 0, \sigma = 1$)

Bibliography

[1] R. E. Bechhofer, S. Elmaghraby and N. Morse, "A Single-Sample Multiple-Decision Procedure for Selecting the Multinomial Event Which Has the Highest Probability," *Annals of Mathematical Statistics,* vol. 30, pp. 102-119, March 1959.

[2] J. O. Miller, B. L. Nelson and C. H. Reilly, "Efficient Multinomial Selection in Simulation," *Naval Research Logistics,* vol. 45, pp. 459-482, 1998.

[3] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin,* vol. 1, no. 6, pp. 80-83, December 1945.

[4] H. B. Mann and D. R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other," *The Annals of Mathematical Statistics,* vol. 18, no. 1, pp. 50-60, March 1947.

[5] S. S. Gupta, "On a decision rule for a problem in ranking means. Doctoral dissertation," *Institute of Statistics, University of North Carolina,* 1956.

[6] S. S. Gupta, "On some multiple decision (ranking and selection) rules," *Technometrics,* vol. 7, pp. 225-245, 1965.

[7] R. E. Bechhofer, "A single-sample multiple decision procedure for ranking means of normal populations with known variances," *Annals of Mathematical Statistics,* vol. 25, pp. 16-39, 1954.

[8] B. L. Nelson and D. Goldsman, "Comparisons with a Standard in Simulation Experiments," *Management Science,* vol. 47, pp. 449-463, 2001.

[9] S. Kim and B. L. Nelson, "Selecting the Best System," in *Handbooks in Operations Research and Management Science: Simulation,* S. G. Henderson and B. L. Nelson, Eds., Oxford: Elsevier Science, 2006.

[10] L. W. Koenig and A. M. Law, "A Procedure for Selecting a Subset of Size m Containing the l Best of k Independent Normal Populations, with Applications to Simulation," *Communications in Statistics - Simulation and Computation,* vol. 14, no. 3, pp. 719-734, 1985.

[11] S. S. Gupta and T. J. Santner, "On Selection and Ranking Procedures - A Restricted Subset Selection Rule," in *Proc. 39th Session of the International Statistical Institute*, Vienna, 1973.

[12] A. M. Law, Simulation Modeling and Analysis, 4th ed., McGraw-Hill, 2007, p. 768.

[13] D. Goldsman, S. Kim and B. L. Nelson, "Statistical selection of the best system," in *2005 Proceedings of the Winter Simulation Conference*, 2005.

[14] A. Wald, Sequential Analysis, New York: John Wiley, 1947.

[15] E. Paulson, "A Sequential Procedure for Selecting the Population with the Largest Mean from k Normal Populations," *The Annals of Mathematical Statistics,* vol. 35, no. 1, pp. 174-180, March 1964.

[16] D. Goldsman, "On Selecting the Best of k Systems; An Expository Survey of Indifference-Zone Multinomial Procedures, Proceedings of the 1984 Winter Simulation Conference," *The Institute of Electrical and Electronics Engineers,* pp. 107-112, 1984.

[17] J. O. Miller, B. L. Nelson and C. H. Reilly, "Estimating the Probability That a Simulated System Will Be the Best," *Naval Research Logistics,* vol. 49, pp. 341-358, 2002.

[18] R. E. Bechhofer and M. Sobel, "Non-Parametric Multiple-Decision Procedures for Selecting That One of k Populations Which Has the Highest Probability of Yielding the Largest Observation," *Annals of Mathematical Statistics,* vol. 29, p. 325, March 1958.

[19] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics,* vol. 27, pp. 832-837, 1956.

[20] E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics,* vol. 33, pp. 1065-1076, 1962.

[21] E. Fix and J. L. Hodges, "Discriminatory analysis - nonparametric discrimination: consistency properties," *Report No. 4, Project no. 21-29-004,* 1951.

[22] H. Akaike, "An approximation to the density function," *Annals of the Institute of Statistical Mathematics,* vol. 6, pp. 127-132, 1954.

[23] M. P. Wand and M. C. Jones, Kernel Smoothing, Boca Raton, FL: Chapman & Hall/CRC, 1995.

[24] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC, 1986, p. 48.

[25] D. D. Wackerly, W. Mendenhall III and R. L. Scheaffer, Mathematical Statistics wtih Applications, 7th ed., Brooks/Cole, Cengage Learning, 2008.

[26] M. Hollander and D. A. Wolfe, Nonparametric Statistical Methods, 2nd ed., New York: John Wiley & Sons, 1999.

[27] R. H. Randles and D. A. Wolfe, Introduction to the Theory of Nonparametric Statistics, Malabar, Florida: Krieger Publishing Company, 1991.

[28] G. Casella and R. L. Berger, Statistical Inference, Pacific Grove, California: Wadsworth and Brooks/Cole, 1990, p. 216.

Vita

Captain Lessin entered the Air Force Academy in June 2004 and was commissioned upon graduating in May 2008 with a Bachelor of Science degree in mathematics. After graduation, he was assigned to the Air Force Operational Test and Evaluation Center (AFOTEC), Detachment 4, at Peterson Air Force Base. He worked as a space force enhancement test analyst, planning and conducting operational tests on the Air Force's newest satellite technologies. In August 2011 he began his studies at the Air Force Institute of Technology in pursuit of a Masters degree in Operations Research. Upon graduation, Captain Lessin will be assigned to the 53rd Test Management Group, Detachment 4, at Creech Air Force Base, where he will work as an Unmanned Aerial Vehicle (UAV) operational test analyst.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 21-03-2013	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From – To)* Oct 2011 – Mar 2013

4. TITLE AND SUBTITLE ESTIMATING THE PROBABILITY OF BEING THE BEST SYSTEM: A GENERALIZED METHOD AND NONPARAMETRIC HYPOTHESIS TEST	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Lessin, Aaron M., Captain, USAF	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB OH 45433-7765	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT-ENS-13-M-10
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) SES USAF AFMC AFSC/LG Attn: Gilbert J. Montoya BLDG 3001 POST 1AG76A Tinker AFB, OK 73145 (405) 736-2863 (DSN: 336-2863) Gilbert.Montoya@tinker.af.mil	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFSC/LG **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

This thesis provides two new approaches for comparing competing systems. Instead of making comparisons based on long run averages or mean performance, the first paper presents a generalized method for calculating the probability that a single system is the best among all systems in a single trial. Unlike current empirical methods, the generalized method calculates the exact multinomial probability that a single system is best among competing systems. The ability to avoid time consuming empirical estimate techniques could potentially result in significant savings in both time and money when comparing alternate systems. A Monte Carlo simulation is conducted comparing the empirical probability estimates of the generalized integral method, calculated using a bootstrapping procedure and density estimation technique, with those of two related estimation techniques, Procedure BEM (Bechhofer, Elmaghraby, and Morse) and Procedure AVC (All Vector Comparisons). All test cases show comparable performance in empirical estimation accuracy of the generalized integral method with that of the current methods analyzed.

The second paper proposes the use of a distribution-free ordered comparisons procedure to test whether one population is stochastically larger (or smaller) than the other. This procedure is suggested as a useful alternative to the well-known Wilcoxon rank sum and Mann-Whitney tests. A Monte Carlo study is conducted, comparing the methods based on simulated power and type I error. All test cases show a marked improvement in performance of the ordered comparisons test versus the Wilcoxon rank sum test, when testing for stochastic dominance. A table of critical values is presented and a large sample approximation is given.

15. SUBJECT TERMS

Wilcoxon rank sum test, ranking and selection, stochastic dominance, simulation output analysis, multinomial probability

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** J.O. Miller, Ph.D. (ENS)
REPORT U	**ABSTRACT** U	**c. THIS PAGE** U	UU	114	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 7469 e-mail: john.miller@afit.edu